

UNITED STATES AIR FORCE
SUMMER RESEARCH PROGRAM -- 1997
SUMMER RESEARCH EXTENSION PROGRAM FINAL REPORTS

VOLUME 3
ROME LABORATORY

RESEARCH & DEVELOPMENT LABORATORIES
5800 Uplander Way
Culver City, CA 90230-6608

Program Director, RDL
Gary Moore

Program Manager, AFOSR
Colonel Jan Cervený

Program Manager, RDL
Scott Licoscós

Program Administrator, RDL
Johnetta Thompson

Program Administrator, RDL
Rebecca Kelly-Clemmons

Submitted to:

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH
Bolling Air Force Base
Washington, D.C.
December 1997

20010319 052

AQM01-06-1189

PREFACE

This volume is part of a five-volume set that summarizes the research of participants in the 1997 AFOSR Summer Research Extension Program (SREP.)

Reports presented in this volume are arranged alphabetically by author and are numbered consecutively – e.g., 1-1, 1-2, 1-3; 2-1, 2-2, 2-3. Reports in the five-volume set are organized as follows:

| VOLUME | TITLE |
|--------|--|
| 1 | Armstrong Laboratory |
| 2 | Phillips Laboratory |
| 3 | Rome Laboratory |
| 4A | Wright Laboratory |
| 4B | Wright Laboratory |
| 5 | Arnold Engineering Development Center Air Logistics Centers United States Air Force Academy Wilford Hall Medical Center |

1997 SREP Final Technical Report Table of Contents

Armstrong Laboratory

Volume 1

| | Principle Investigator | Report Title University/Institution | Laboratory & Directorate |
|----|----------------------------|--|--------------------------|
| 1 | Dr. Richelle M. Allen-King | Trans-1,2-Dichloroethene Transformation Rate in a Metallic Iron/Water System: Effects of Concentration and Temperature Washington State University | AL/EQC |
| 2 | Dr. Anthony R. Andrews | Development of Multianalyte Electrochemiluminescence Sensors & Biosensors Ohio University | AL/EQC |
| 3 | Dr. Jer-Sen Chen | Development of Perception Based Video Compression Algorithms Using Reconfigurable Hardware Wright State University | AL/CFHV |
| 4 | Dr. Cheng Cheng | Investigation & Eval of Optimization Algorithms Guiding the Assignment of Recruits to Training School Seats John Hopkins University | AL/HRM |
| 5 | Dr. Randolph D. Glickman | Optical Detection of Intracellular Photooxidative Reactions University of Texas Health Science Center | AL/OEO |
| 6 | Dr. Nandini Kannan | Predicting Altitude Decompression sickness Using Survival Models University of Texas at San Antonio | AL/CFTS |
| 7 | Dr. Antti J. Koivo | Skill Improvements Via Reflected Force Feedback Purdue Research Foundation | AL/CFBA |
| 8 | Dr. Suk B. Kong | Degradation & Toxicology Studies of JP-8 Fuel in Air, Soil & Drinking Water Incarnate Word College | AL/OEA |
| 9 | Dr. Audrey D. Levine | Biogeochemical Assessment of Natl Attenuation of JP-4 Contaminated Ground in the Presence of Fluorinated Surfactants Utah State University | AL/EQC |
| 10 | Dr. Robert G. Main | The Effect of Video Image Size & Screen Refresher Rate On Mess Retention Cal State University, Chico | AL/HRT |
| 11 | Dr. Phillip H. Marshall | On the Resilience of Time-to-Contact Judgements: The Determination of Inhibitory and Facilitory Influences, and Factor Structure Texas Tech University | AL/HRM |
| 12 | Dr. Bruce V. Mutter | Environmental cost Analysis: Calculating Return on Investment for Emerging Technologies Bluefield State College | AL/EQP |

REPORT DOCUMENTATION PAGE

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering the data, reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing the burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Project (0770-0188).

AFRL-SR-BL-TR-00-

0770

| | | | | | |
|---|--|---|----------------------------------|--|--|
| 1. AGENCY USE ONLY (Leave blank) | | 2. REPORT DATE December, 1997 | | 3. REPORT TYPE | |
| 4. TITLE AND SUBTITLE 1997 Summer Research Program (SRP), Summer Research Extension Program (SREP), Final Report, Volume 3, Rome Laboratory | | | | 5. FUNDING NUMBERS F49620-93-C-0063 | |
| 6. AUTHOR(S) Gary Moore | | | | | |
| 7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Research & Development Laboratories (RDL) 5800 Uplander Way Culver City, CA 90230-6608 | | | | 8. PERFORMING ORGANIZATION REPORT NUMBER | |
| 9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) Air Force Office of Scientific Research (AFOSR) 801 N. Randolph St. Arlington, VA 22203-1977 | | | | 10. SPONSORING/MONITORING AGENCY REPORT NUMBER | |
| 11. SUPPLEMENTARY NOTES | | | | | |
| 12a. DISTRIBUTION AVAILABILITY STATEMENT Approved for Public Release | | | | 12b. DISTRIBUTION CODE | |
| 13. ABSTRACT (Maximum 200 words) The United States Air Force Summer Research Program (SRP) is designed to introduce university, college, and technical institute faculty members to Air Force research. This is accomplished by the faculty members, graduate students, and high school students being selected on a nationally advertised competitive basis during the summer intersession period to perform research at Air Force Research Laboratory (AFRL) Technical Directorates and Air Force Air Logistics Centers (ALC). AFOSR also offers its research associates (faculty only) an opportunity, under the Summer Research Extension Program (SREP), to continue their AFOSR-sponsored research at their home institutions through the award of research grants. This volume consists of a listing of the participants for the SREP and the technical report from each participant working at the AF Rome Laboratory. | | | | | |
| 14. SUBJECT TERMS Air Force Research, Air Force, Engineering, Laboratories, Reports, Summer, Universities, Faculty, Graduate Student, High School Student | | | | 15. NUMBER OF PAGES | |
| | | | | 16. PRICE CODE | |
| 17. SECURITY CLASSIFICATION OF REPORT Unclassified | 18. SECURITY CLASSIFICATION OF THIS PAGE Unclassified | 19. SECURITY CLASSIFICATION OF ABSTRACT Unclassified | 20. LIMITATION OF ABSTRACT UL | | |

GENERAL INSTRUCTIONS FOR COMPLETING SF 298

The Report Documentation Page (RDP) is used in announcing and cataloging reports. It is important that this information be consistent with the rest of the report, particularly the cover and title page. Instructions for filling in each block of the form follow. It is important to **stay within the lines** to meet **optical scanning requirements**.

Block 1. Agency Use Only (Leave blank).

Block 2. Report Date. Full publication date including day, month, and year, if available (e.g. 1 Jan 88). Must cite at least the year.

Block 3. Type of Report and Dates Covered. State whether report is interim, final, etc. If applicable, enter inclusive report dates (e.g. 10 Jun 87 - 30 Jun 88).

Block 4. Title and Subtitle. A title is taken from the part of the report that provides the most meaningful and complete information. When a report is prepared in more than one volume, repeat the primary title, add volume number, and include subtitle for the specific volume. On classified documents enter the title classification in parentheses.

Block 5. Funding Numbers. To include contract and grant numbers; may include program element number(s), project number(s), task number(s), and work unit number(s). Use the following labels:

C - Contract
G - Grant
PE - Program
Element

PR - Project
TA - Task
WU - Work Unit
Accession No.

Block 6. Author(s). Name(s) of person(s) responsible for writing the report, performing the research, or credited with the content of the report. If editor or compiler, this should follow the name(s).

Block 7. Performing Organization Name(s) and Address(es). Self-explanatory.

Block 8. Performing Organization Report Number. Enter the unique alphanumeric report number(s) assigned by the organization performing the report.

Block 9. Sponsoring/Monitoring Agency Name(s) and Address(es). Self-explanatory.

Block 10. Sponsoring/Monitoring Agency Report Number. (// known)

Block 11. Supplementary Notes. Enter information not included elsewhere such as: Prepared in cooperation with....; Trans. of....; To be published in.... When a report is revised, include a statement whether the new report supersedes or supplements the older report.

Block 12a. Distribution/Availability Statement. Denotes public availability or limitations. Cite any availability to the public. Enter additional limitations or special markings in all capitals (e.g. NOFORN, REL, ITAR).

DOD - See DoDD 5230.24, "Distribution Statements on Technical Documents."

DOE - See authorities.

NASA - See Handbook NHB 2200.2.

NTIS - Leave blank.

Block 12b. Distribution Code.

DOD - Leave blank.

DOE - Enter DOE distribution categories from the Standard Distribution for Unclassified Scientific and Technical Reports.
Leave blank.

NASA - Leave blank.

NTIS -

Block 13. Abstract. Include a brief (*Maximum 200 words*) factual summary of the most significant information contained in the report.

Block 14. Subject Terms. Keywords or phrases identifying major subjects in the report.

Block 15. Number of Pages. Enter the total number of pages.

Block 16. Price Code. Enter appropriate price code (*NTIS only*).

Blocks 17. - 19. Security Classifications. Self-explanatory. Enter U.S. Security Classification in accordance with U.S. Security Regulations (i.e., UNCLASSIFIED). If form contains classified information, stamp classification on the top and bottom of the page.

Block 20. Limitation of Abstract. This block must be completed to assign a limitation to the abstract. Enter either UL (unlimited) or SAR (same as report). An entry in this block is necessary if the abstract is to be limited. If blank, the abstract is assumed to be unlimited.

1997 SREP Final Technical Report Table of Contents

Armstrong Laboratory

Volume 1 (cont.)

| | Principle Investigator | Report Title University/Institution | Laboratory & Directorate |
|----|------------------------|--|--------------------------|
| 13 | Dr. Sundaram Narayanan | Java-Based Interactive Simulation Architecture for Airbase Logistics Modeling Wright State University | AL/HRT |
| 14 | Dr. Barth F. Smets | Coupling of 2, 4-&2, 6-Dinitrotoluene Mineralization W/NO ₂ Removal by University of Cincinnati | AL/EQC |
| 15 | Dr. Mary Alice Smith | In Vitro Detection of Apoptosis in Differentiating Mesenchymal Cells Using Immunohistochemistry and Image Analysis University of Georgia | AL/OET |
| 16 | Dr. William A. Stock | Application of Meta-Analysis to Research on Pilot Training: Extensions to Flight Simulator Visual System Research Arizona State University | AL/HRA |
| 17 | Dr. Nancy J. Stone | Evaluation of a Scale Designed to Measure the Underlying Constructs of Engagement, Involvement, & Self-Regulated Learning Creighton University | AL/HRT |
| 18 | Dr. Mariusz Ziejewski | Characterization of Human Head/Neck Response in Z-Direction in Terms of Significant Anthropomorphic Parameters, Gender, Helmet Weight and Helmet Center North Dakota State University | AL/CFBV |
| 19 | Dr. Kevin M. Lambert | Magnetic Effects on the Deposition & Dissolution of Calcium Carbonate Scale Brigham Young University | AL/EQS |
| 20 | Dr. Jacqueline C. Shin | Coordination of Cognitive & Perceptual-Motor Activities Pennsylvania State University | AL/HRM |
| 21 | Dr. Travis C. Tubre | The Development of a General Measure of Performance Texas A&M University-College Station | AL/HRT |
| 22 | Dr. Robert B. Trelease | Development of Qualitative Process Modeling Systems for Cytokines, Cell Adhesion Molecules, and Gene Regulation University of California – Los Angeles | AL/AOH |

1997 SREP Final Technical Report Table of Contents

Phillips Laboratory

Volume 2

| | Principle Investigator | Report Title University/Institution | Laboratory & Directorate |
|----|----------------------------|---|--------------------------|
| 1 | Dr. Graham R. Allan | Temporal & Spatial Characterization of a Synchronously-Pumped New Mexico Highlands University | PL/LIDN |
| 2 | Dr. Joseph M. Calo | Transient Studies of the Effects of Fire Suppressants in a Well-Stirred Combustor Brown University | PL/GPID |
| 3 | Dr. James J. Carroll | Examination of Critical Issues in the Triggering of Gamma Rays from 178Hfm2 Youngstown State University | PL/WSQ |
| 4 | Dr. Soyoung S. Cha | Gradient-Data Tomography for Hartman Sensor Application to Aero-Optical Field Reconstruction University of Illinois at Chicago | PL/LIMS |
| 5 | Dr. Judith E. Dayhoff | Dynamic Neural Networks: Towards Control of Optical Air Flow Distortions University of Maryland | PL/LIMS |
| 6 | Dr. Ronald R. DeLyser | Computational Evaluation of Optical Sensors University of Denver | PL/WSTS |
| 7 | Dr. Andrew G. Detwiler | Analysis & Interpretation of Contrail Formation Theory & Observations South Dakota School of Mines – Technology | PL/GPAB |
| 8 | Dr. Itzhak Dotan | Measurements of Ion-Molecule Reactions at Very High Temperature The Open University of Israel | PL/GPID |
| 9 | Dr. George W. Hanson | Electromagnetic Modeling of Complex Dielectric/Metallic Mines In A Layered University of Wisconsin – Milwaukee | PL/WSQ |
| 10 | Dr. Mayer Humi | Optical & Clear Air Turbulence Worcester Polytechnic Inst. | PL/GPAA |
| 11 | Dr. Christopher H. Jenkins | Shape Control of an Inflated Circular Disk Experimental Investigation South Dakota School of Mines – Technology | PL/VT |
| 12 | Dr. Dikshitulu K. Kalluri | Numerical Simulation of Electromagnetic Wave Transformation in a Dynamic Magnetized Plasma University of Lowell | PL/GPIA |
| 13 | Dr. Aravinda Kar | Improved Chemical Oxygen-Iodine Laser (COIL) Cutting Models to Optimize Laser Parameters University of Central Florida | PL/LIDB |

1997 SREP Final Technical Report Table of Contents

Phillips Laboratory

Volume 2 (cont.)

| | Principle Investigator | Report Title University/Institution | Laboratory & Directorate |
|----|--------------------------|--|--------------------------|
| 14 | Dr. Andre Y. Lee | Characterization of Thermoplastic Inorganic-Organic Hybrid Polymers Michigan State University | PL/RKS |
| 15 | Dr. Feng-Bao Lin | Improvement in Fracture Propagation Modeling for Structural Ballistic Risk Assessment Polytechnic University of New York | PL/RKEM |
| 16 | Dr. Ronald A. Madler | Cross Sectional Area Estimation of Orbital Debris Embry-Riddle Aeronautical University | PL/WSAT |
| 17 | Dr. Carlos A. Ordonez | Incorporation of Boundary condition Models into the AF Computer Simulation University of North Texas | PL/WSQA |
| 18 | Dr. James M. Stiles | Wide Swath, High Resolution, Low Ambiguity SAR Using Digital Beamforming Arrays University of Kansas | PL/VTRA |
| 19 | Dr. Charles M. Swenson | Balloon Retromodulator Experiment Post- flight Analysis Utah State University | PL/VTRA |
| 20 | Dr. Miguel Velez-Reyes | Development of Algorithms for Linear & Nonlinear Retrieval Problems in Atmospheric Remote Sensing University of Puerto Rico | PL/GPAS |
| 21 | Dr. John D. Holtzclaw | Experimental Investigation of Ipinging Jets University of Cincinnati | PL/RKS |
| 22 | Dr. Jeffrey W. Nicholson | Radar Waves with Optical Carriers University of New Mexico | PL/LIDB |

1997 SREP Final Technical Report Table of Contents

Rome Laboratory

Volume 3

| Principle Investigator | Report Title University/Institution | Laboratory & Directorate |
|------------------------------|---|--------------------------|
| 1 Dr. A. F. Anwar | Deep Quantum Well Channels for Ultra Low Noise HEMTs for Millimeter and Sub-millimeter Wave Applications University of Connecticut | AFRL/SNH |
| 2 Dr. Ahmed E. Barbour | Investigating the Algorithmic Nature of the Proof Structure of ORA Larch/VHDL Georgia Southern University | RL/ERDD |
| 3 Dr. Milica Barjaktarovic | Specification & Verification of MISSI Architecture Using SPIN Wilkes University | RL/C3AB |
| 4 Dr. Daniel C. Bukofzer | Analysis, Performance Evaluations, & Computer Simulations of Receivers Processing Low Probability of Intercept Signals Cal State Univ. Fresno | RL/C3BA |
| 5 Dr. Xuesheng Chen | Non-Destructive Optical Characterization of Composition & Its Uniformity in Multilayer Ternary Semiconductor Stacks Wheaton College | RL/ERX |
| 6 Dr. Jun Chen | Amplitude Modulation Using Feedback Sustained Pulsation as Sub-Carrier in Rochester Inst of Technol | RL/OCPA |
| 7 Dr. Everett E. Crisman | Development of Anti-Reflection Thin Films for Improved Coupling of Laser Energy into Light Activated, Semiconductor Re-Configurable, Microwave Source/Antenna Brown University | RL/ERAC |
| 8 Dr. Digendra K. Das | Development of a Stimulation Model for Determining the Precision Of Reliability SUNYIT | RL/ERSR |
| 9 Dr. Matthew E. Edwards | An Application of PROFILER for Modeling the Diffusion of Of Aluminum-Copper on a Silicon Substrate Spelman College | RL/ERDR |
| 10 Dr. Kaliappan Gopalan | Analysis of Stressed Speech Using Cepstral Domain Features Purdue University – Calumet | RL/IRAA |
| 11 Dr. James P. LeBlanc | Multichannel Autoregressive Modeling & Multichannel Innovations Based New Mexico State University | RL/OCSS |
| 12 Dr. Hrushikesh N. Mhaskar | Multi-Source Direction Finding Cal State University, Los Angeles | RL/ERAA |
| 13 Dr. Ronald W. Noel | An Evolutionary Sys for Machine Recognition of Software Source Code Rensselaer Polytechnic Inst | RL/C3CA |

1997 SREP Final Technical Report Table of Contents

Rome Laboratory

Volume 3 (cont.)

| Principle Investigator | Report Title University/Institution | Laboratory & Directorate |
|--------------------------|--|--------------------------|
| 14 Dr. Glenn E. Prescott | Rapid Prototyping of Software Radio Sys Using Field Programmable Gate Arrays University of Kansas Center for Research | RL/C3BB |
| 15 Dr. Mysore R. Rao | Wavelet Transform Techniques for Isolation, Detection & Classification of Concealed Objects in Images Rochester Institute of Technology | RL/OCSM |
| 16 Dr. Scott E. Spetka | IPL HTML Interface Performance Evaluation SUNY of Tech Utica | RL/IRD |
| 17 Dr. Gang Sun | Investigation of Si/ZnS Near Infrared Intersubband Lasers University of Massachusetts-Boston | RL/EROC |
| 18 Mr. Parker E. Bradley | Development of a User-Friendly Computer Environment for Blind Source Syracuse University | RL/C3BB |

1997 SREP Final Technical Report Table of Contents

Wright Laboratory

Volume 4A

| Principle Investigator | Report Title University/Institution | Laboratory & Directorate |
|----------------------------|--|--------------------------|
| 1 Dr. Mohammad S. Alam | Infrared Image Registration & High Resolution Reconstruction Using Rotationally Translated Video Sequences* Purdue University | WL/AAJT |
| 2 Dr. Pnina Ari-Gur | Optimizing Microstructure, Texture & Orientation Image Microscopy of Hot Rolled Ti-6Al-4V Western Michigan University | WL/MLLN |
| 3 Dr. James D. Baldwin | Multi-Site & Widespread Fatigue Damage in Aircraft Structure in the Presence of Prior Corrosion University of Oklahoma | WL/FIB |
| 4 Dr. Armando R. Barreto | Deconvolution of the Space-Time Radar Spectrum Florida International University | WL/AAMR |
| 5 Dr. Marc M. Cahay | Improved Modeling of Space-Charge Effects in a New Cold Cathode Emitter University of Cincinnati | WL/AADM |
| 6 Dr. Reaz A. Chaudhuri | Interfacing of Local Asymptotic Singular & Global Axisymmetric Micromechanical University of Utah | WL/MLBM |
| 7 Dr. Robert J. DeAngelis | Texture Formation During the Thermo-Mechanical Processing of Copper Plate University of Nebraska – Lincoln | WL/MNMW |
| 8 Dr. Gregory S. Elliott | The Study of a Transverse Jet in a Supersonic Cross-Flow Using Advanced Laser Rutgers: State University of New Jersey | WL/POPT |
| 9 Dr. Altan M. Ferendeci | Development of Multiple Metal-Dielectric Layers for 3-D MMIC University of Cincinnati | WL/AADI |
| 10 Dr. Allen G. Greenwood | Development of a Prototype to Test & Demonstrate the MODDCE Framework Mississippi State University | WL/MTI |
| 11 Dr. Michael A. Grinfeld | Mismatch Stresses & Lamellar Microstructure of TiAl-Alloys Rutgers University- Piscataway | WL/MLLM |
| 12 Dr. Michael C. Larson | Interfacial Sliding in Brittle Fibrous Composites Tulane University | WL/MLLM |
| 13 Dr. Douglas A. Lawrence | Tools for the Analysis & Design of Gain Scheduled Missile Autopilots Ohio University | WL/MNAG |

1997 SREP Final Technical Report Table of Contents

Wright Laboratory (cont.)

Volume 4A

| | Principle Investigator | Report Title University/Institution | Laboratory & Directorate |
|----|-------------------------|---|--------------------------|
| 14 | Dr. Junghsen Lieh | Determination of 3D Deformations, Forces & Moments of Aircraft Wright State University | WL/FIVM |
| 15 | Dr. Zongli Lin | Control of Linear Sys w/Rate Limited Actuators & Its Applications to Flight Control Systems SUNY Stony Brook | WL/FI |
| 16 | Dr. Paul Marshall | Experimental & Computational Investigations of Bromine & Iodine Chemistry in Flame Suppression University of North Texas | WL/MLBT |
| 17 | Dr. Hui Meng | Development of Holographic Visualization & Holographic Velocimetry Techniques Kansas State University | WL/POSC |
| 18 | Dr. Douglas J. Miller | Band Gap Calculations on Squarate-Containing Conjugated Oligomers for the Prediction of Conductive and Non-Linear Optical Properties of Polymeric Materials Cedarville College | WL/MLBP |
| 19 | Dr. Timothy S. Newman | Classification & Visualization of Tissue in Multiple Modalities of Brain MR University of Alabama at Huntsville | WL/AACR |
| 20 | Dr. Mohammed Y. Niamat | FPGA Implementation of the Xpatch Ray Tracer University of Toledo | WL/AAST |
| 21 | Dr. Anthony C. Okafor | Development of Optimum Drilling Process for Advanced Composites University of Missouri – Rolla | WL/MTI |
| 22 | Dr. George A. Petersson | Absolute Rates for Chemical Reactions Wesleyan University | WL/MLBT |
| 23 | Dr. Mohamed N. Rahaman | Process Modeling of the Densification of Granular Ceramics Interaction Between Densification and Creep University of Missouri – Rolla | WL/MLLN |

1997 SREP Final Technical Report Table of Contents

Wright Laboratory (cont.)

Volume 4B

| | Principle Investigator | Report Title University/Institution | Laboratory & Directorate |
|----|-------------------------|--|--------------------------|
| 24 | Dr. Martin Schwartz | Quantum Mechanical Modeling of the Thermochemistry of Halogenated Fire Suppressants University of North Texas | WL/MLBT |
| 25 | Dr. Marek Skowronski | Investigation of Slip Boundaries in 4H-SiC Crystals Carnegie Mellon University | WL/MLPO |
| 26 | Dr. Yong D. Song | Guidance & Control of Missile Sys Under Uncertain Flight Conditions North Carolina A&T State University | WL/MNAG |
| 27 | Dr. Raghavan Srinivasan | Models for Microstructural Evolution During Dynamic Recovery Wright State University | WL/MLIM |
| 28 | Dr. Scott K. Thomas | The Effects of Transient Acceleration Loadings on the Performance of a Copper-Ethanol Heat Pipe with Spiral Grooves Wright State University | WL/POOS |
| 29 | Dr. James P. Thomas | The Effect of Temperature on Fatigue Crack Growth of TI-6AL-4V in the Ripple University of Notre Dame | WL/MLLN |
| 30 | Dr. Karen A. Tomko | Scalable Parallel Solution of the 3D Navier-Stokes Equations Wright State University | WL/FIM |
| 31 | Dr. J. M. Wolff | Off Design Inviscid/Viscous Forced Response Prediction Model for High Cycle Wright State University | WL/POTF |
| 32 | Mr. Todd C. Hathaway | Experiments on Consolidation of Aluminum Powders Using Simple Shear University of North Texas | WL/MLLN |
| 33 | Ms. Diana M. Hayes | Error Correction & Compensation for Mueller Matrices Accounting for Imperfect Polarizers University of North Texas | WL/MNGA |

1997 SREP Final Technical Report Table of Contents

Volume 5

| | Principle Investigator | Report Title University/Institution | Laboratory & Directorate |
|--|---------------------------|--|-----------------------------|
| Arnold Engineering Development Center | | | |
| 1 | Dr. Frank G. Collins | Development of Laser Vapor Screen Flow Visualization Sys Tennessee University Space Institute | AEDC |
| United States Air Force Academy | | | |
| 2 | Mr. Derek E. Lang | Experimental Investigation of Liquid Crystal Applications for Boundary Layer Characterization University of Washington | USAFA/DFA |
| Air Logistics Centers | | | |
| 3 | Dr. Sandra A. Ashford | Development of Jet Engine Test Facility Vibration Signature & Diagnostic System University of Detroit Mercy | OCALC/TIE |
| 4 | Dr. Roger G. Ford | Use of Statistical Process Control in a Repair/Refurbish/ Remanufactureg Environment St. Mary's University | SAALC |
| Wilford Hall Medical Center | | | |
| 5 | Dr. Stedra L. Stillmana | Metabolite Profile Following the Administration of Fenproporex University of Alabama at Birmingham | WHMC |

Deep Quantum Well Channels for Ultra Low Noise HEMTs for Millimeter and Sub-millimeter Wave Applications

A. F. M. Anwar

Associate Professor

Electrical and Systems Engineering Department

The University of Connecticut

Storrs, CT 06269-2157

Final Report for:

Summer Research Extension Program

Sponsored by:

Air Force of Scientific Research

Bolling Air Force Base, DC

and

Rome Laboratory

Hanscom Air Force Base

January 1998

Deep Quantum Well Channels for Ultra Low Noise HEMTs for Millimeter and Sub-millimeter Wave Applications

A. F. M. Anwar

Associate Professor

Electrical and Systems Engineering Department

The University of Connecticut

Storrs, CT 06269-2157

Abstract

A complete model that includes (a) quantum well calculations (b) transport and (c) noise for deep quantum well HEMT is presented. Schroedinger and Poisson's equations are solved self-consistently to determine the quantum well parameters of the deep quantum well formed in AlGaAsSb/InGaAs/AlGaAsSb. Time independent Boltzman transport equation (BTE) is solved using ensemble Monte Carlo to study the velocity-electric field characteristic in $\text{In}_x\text{Ga}_{1-x}\text{As}$. It is seen that impact ionization in narrow band gap channel gives rise to a degradation in the low field mobility and peak velocity. The effect of impact ionization is modeled by a voltage dependent current source in the small signal equivalent circuit. The current source is characterized by an impact ionization induced transconductance whose magnitude can be extracted from the experimental Y-parameter data. The solution of the time dependent BTE indicates that impact ionization process is frequency dependent and this dependence is also a function of drain current and source drain spacing. $\text{In}_x\text{Ga}_{1-x}\text{As}_y\text{Sb}_{1-y}$ provides excellent transport properties and may be used as a channel material in HEMTs. The usefulness of the quaternary channel can be further enhanced if a superlattice channel is used instead of an alloyed channel.

Deep Quantum Well Channels for Ultra Low Noise HEMTs for Millimeter and Sub-millimeter Wave Applications

I. Introduction

A model for study deep QW channel HEMTs is presented. The theoretical calculations include a) self consistent solution of Schroedinger and Poisson's equation and b) transport in InGaAs ternary channel. Transport is investigated by solving Boltzman Transport Equation (BTE) using ensemble Monte Carlo simulation. The solution of Schroedinger and Poisson's along with the description of the device along the growth direction yields charge control. Transport data, such as low field mobility and saturation velocity given the device layout and charge control provides the d.c. I-V curves and the small signal parameters. The availability of the small signal parameters enable one to model noise and the device performance at high frequency.

In Sec.2, the method used to solve Schroedinger and Poisson's is discussed along with relevant results. In Sec.3, the transport in the channel material is presented. The effect of impact ionization is taken into account in the transport study.. In Sec.4, the modeled I-V and small signal parameters are presented. The calculation of drain current takes into effect impact ionization in the channel. In Sec.5, noise in the channel in devices is addressed theoretically. In Sec.6, the frequency spectrum of impact ionization is investigated by solving time dependent BTE. The use of InGaAsSb as the ideal channel material is investigated in Sec. 7. A conclusion is provided in Sec. 8.

II.1 Quantum well and material characterization

II.1.1 Material Data

The calculation of conduction band discontinuity ΔE_c in AlGaAsSb/InAs is reported by Anwar and Webster¹ and is an extension of Schuermeyer et. al.². A diagram is constructed beginning with known valence band energy differences for binary systems and then adding the bandgap energies to find the conduction band minimum for each binary compound. The energy bandgaps of the ternary systems are computed by interpolation over alloy composition using bowing parameters¹. The calculation of the energy bandgap and the lattice constants of the quaternary follows the work of Moon et. al.³, Glisson et. al.⁴ and Svensson et. al.⁵ The energy bandgap of the quaternary can be written as:

$$E_G^Q(x,y) = (1-x)T_{14}(y) + xT_{23}(y) - \Delta(x,y) \quad (\text{II.1-1})$$

where $T_{ij}(y) = yB_j + (1-y)B_i - y(1-y)C_{ij}$ are the ternary alloy bandgaps, B_i 's are the bandgap of the binaries, C_{ij} are the bowing parameters for the ternary alloy, and $\Delta(x,y) = x(1-x)[(1-y)C_{12} + yC_{43}] + x(1-x)y(1-y)C_Q$, where $C_Q = C_{14}x + C_{23}(1-x)$. Using the parameters listed in Ref.¹, $E_G^Q(x,y)$ is calculated for the Γ -, X-, and L-valleys and the lowest result is chosen as the bandgap. The lattice constant of the quaternary L_Q is interpolated using the following relationship:

$$L_Q = L_1 + (L_2 - L_1)x + (L_4 - L_1)y + (L_1 - L_2 + L_3 - L_4)xy \quad (\text{II.1-2})$$

where L_i 's are the lattice constants of the binaries¹. Finally, ΔE_c for the lattice matched $\text{Al}_x\text{Ga}_{1-x}\text{As}_{1-y}\text{Sb}_y/\text{InAs}$ is determined by calculating the difference between the conduction band energies of InAs and $\text{Al}_x\text{Ga}_{1-x}\text{As}_{1-y}\text{Sb}_y$.

II.1.2 An Envelope Function Description of Deep Quantum Wells

In this section we report a self-consistent solution to model the QW formed in the conduction band of an $\text{AlGaAsSb}/\text{InAs}/\text{AlGaAsSb}$ heterostructure⁶. The results of this analysis will enable us to compute the charge control, current-voltage and noise performance of this class of devices. Furthermore, this analysis will guide material, device and fabrication research to achieve ultra low noise, very high frequency HEMTs.

II.1.3 Mathematical Model

In $\text{AlGaAsSb}/\text{InAs}/\text{AlGaAsSb}$ systems the QW is formed in InAs . The one electron Schrödinger equation, under effective mass approximation, can be written as⁶

$$-\left(\frac{\hbar^2}{2}\right) \frac{d}{dx} \left(\frac{1}{m^*} \frac{d}{dx} \right) \xi_i + (V(x) - E_i) \xi_i = 0 \quad (\text{II.1-3})$$

where m^* is the electron effective mass, \hbar is the reduced Planck's constant, $\xi_i(x)$ is the envelope wave function, E_i is the energy eigen value, $V(x)$ is the potential energy and the subscript "i" denotes the i^{th} subband. For simplicity the potential energy functions approximated by three straight lines with slopes a_1 , a_2 and a_3 , respectively, and is expressed as

$$V(x) = \begin{cases} V_0 & x < 0 \\ a_j x + \Delta E_j & x_{j-1} < x < x_j \quad j=1,2,3 \end{cases} \quad (\text{II.1-4})$$

where $\Delta E_1 = 0$, $\Delta E_2 = (a_1 - a_2)x_1$, $\Delta E_3 = \Delta E_2 + \Delta E_{c2} + L(a_2 - a_3)$, ΔE_{c2} is the conduction band discontinuity at the second heterointerface, $L = x_2 - x_0$ is the width of the well, $x_0 = 0$ is the position of the first heterointerface, and x_3 is the distance from the first heterointerface in which 99% of the electrons reside. The solution to the Schroedinger equation, for different regions, may be written as

$$\begin{aligned} \xi_j(x) &= \alpha_{10} e^{\beta x} + \alpha_{2,0} e^{-\beta x} & j=0 \\ \xi_j(x) &= \alpha_{1,j} \text{Ai}(\zeta_j) + \alpha_{2,j} \text{Bi}(\zeta_j) & j=1,2,3 \end{aligned} \quad (\text{II.1-5})$$

where $\beta = \sqrt{(2m_0^*/\hbar)(\Delta E_{c1} - E)}$, ΔE_{c1} is the conduction band discontinuity at the first heterointerface, m_0^* the electron effective mass in AlGaAsSb, Ai and Bi are the Airy and the complementary function, respectively, $\zeta_j = \gamma_j \left(x + \frac{\Delta E_j - E}{a_j} \right)$, with $\gamma_j = \left(2m_j^* a_j / \hbar^2 \right)^{1/3}$ and $\alpha_{k,j}$'s ($k=1,2$ & $j = 0,1,2,3$) are the arbitrary constants. Here the subscript j refers to region j and the superscript i , whenever used, will refer to the i^{th} subband. The eigen values and eigen functions are determined by applying the two boundary conditions at any interface (a) continuity of the wave function and (b) continuity of the first derivative by taking into account the proper effective mass. Having formulated the Schrödinger equation Poisson's equation is formulated:

$$\epsilon \frac{d^2 \phi}{dx^2} = q \sum_i n_{si} \xi_i^2(x) + q N_a \quad (\text{II.1-6})$$

where $n_{si} = \frac{m_{\text{InAs}}^* kT}{\pi \hbar^2} \ln \left(1 + e^{(E_F - E_i)/kT} \right)$ is the number of electrons per unit area, ζ is the envelope wave function in the i^{th} subband, N_a is the acceptor density in the unintentionally doped AlGaAsSb layer, m_{InAs}^* refers to the effective mass in the channel, T is the temperature and E_F is the Fermi level at the interface relative to the conduction band in the channel at $x=0$. In these equations we have chosen the potential energy at the interface as the reference. The Fermi level E_F is expressed as $E_F = q[\phi(0) - \phi(W)] + E_{F0} + \Delta E_{c2}$, where $\phi(0) - \phi(W)$ is the total band bending in the AlGaAsSb

layer $\phi(0) = 0$, W is the depletion depth, $E_{F0} = -\left[\frac{E_g(T)}{2} + kT \ln \frac{N_a}{n_i(T)}\right]$ is the position of the Fermi level with respect to the conduction band in the bulk AlGaAsSb and $n_i(T)$ is the intrinsic carrier concentration in AlGaAsSb. The slopes a_j of the straight lines, which approximate the shape of the QW, are proportional to the average electric field determined by Poisson's equation. By integrating Poisson's equation twice with respect to x , the slopes can be expressed in the form

$$a_j = (q^2/\epsilon)(f_j n_s + N_a W), \quad j=1,2,3 \quad (\text{II.1-7})$$

where

$$f_j = 1 - \sum_i \frac{n_{si}}{n_s} \frac{1}{x_j - x_{j-1}} \int_{x_{j-1}}^x dx \int_{-\infty}^x \zeta_i^2(x') dx', \quad j=1,2,3 \quad (\text{II.1-7.1})$$

and $n_s = \sum_i n_{si}$ is the channel electron density in cm^{-2} . By solving the one electron Schrödinger equation for the given potential we can obtain the eigen energies and the wave functions for the system. The eigen energies and the wave functions determine the shape of the electron distribution in the quantum well which is then used to solve Poisson's equation. The two equations are solved self-consistently until we have accounted for 99% of the carriers in the quantum well.

II.1.4 Results and Discussion

An $\text{In}_{0.53}\text{Ga}_{0.47}\text{As}/\text{Al}_{0.6}\text{Ga}_{0.4}\text{AsSb}$ double barrier HEMT with a QW width of 100\AA is considered. The calculated band offset is 0.984eV . A low field mobility of $53,000\text{ cm}^2/\text{V-s}$ and a saturation velocity of $3.2 \times 10^7\text{ cm/s}$ is used in the simulation⁷. The donor doping density of $2 \times 10^{18}\text{ cm}^{-3}$ with a doped epilayer thickness of 350\AA is used.

In Fig. 1, the Fermi level and X_{av} are plotted as a function of the 2DEG concentration. The plots are obtained by solving Schroedinger and Poisson's equations self-consistently. The functional forms obtained for E_F and X_{av} . For d.c., small signal and noise analysis E_F and X_{av} are expressed in functional forms as is evident in Secs. 4 and 5.

III Transport in InGaAs

In this section, we mainly discuss the transport properties of $\text{In}_x\text{Ga}_{1-x}\text{As}$. Monte Carlo technique to solve BTE is used very successfully to simulate electrons transport in semiconductors. The Monte Carlo simulation includes the different scattering mechanisms such as acoustic phonon

scattering, polar optical phonon scattering, deformation potential scattering, impurity scattering, alloy scattering, piezoelectric scattering and impact-ionization. This section is concluded with Monte Carlo simulation results.

III.1 Monte Carlo Method

Many parameters of a physical system are governed by probability distributions. Therefore, if a mathematically random distribution is used to model such distributions we can, in principle, generate the physical values of these parameters. This is, broadly speaking, the Monte Carlo method. In practice, of course, the physical distribution may be quite complex and difficult to manipulate even with a computer. The manipulations can be simplified by “mapping” the complex distributions on to a simple pseudo-random distribution; the most convenient pseudo-random distribution is the uniform distribution, which is readily available on most computer system.⁸⁻¹¹

In general, if $p(\phi)$ and $p(r)$ are the respective probability densities, associated with ϕ in the Physical distribution and r in the pseudo-random distribution, then

$$\int_0^1 p(\phi') d\phi' = \int_0^1 p(r') dr' \quad (\text{III.1-1})$$

In a uniform distribution $p(r)=1$ so that (III-1) becomes

$$r = \int p(\phi') d\phi \quad (\text{III.1-2})$$

Hence, provided that this integral can be evaluated in simple closed analytical form, inversion will yield a random value for the physical variable ϕ in terms of the uniformly distributed random number r . A simple example of this technique is the generation of the random flight time of a classical particle in a gas for the case when Γ , the total scattering rate, is constant. The probability of this particle traveling unimpeded for time t and then being scattered at the end of this flight is

$$p(t) = \Gamma e^{-\Gamma t} \quad (\text{III.1-3})$$

From equation (III.1-2)

$$r = \int_0^1 \Gamma e^{-\Gamma t'} dt' = 1 - e^{-\Gamma t} \quad (\text{III.1-4})$$

so that the random flight times are given by

$$t = -\frac{1}{\Gamma} \ln(1-r) = -\frac{1}{\Gamma} \ln(r) \quad (\text{III.1-5})$$

Note that, since r is uniformly distribution, $\ln(1-r)$ is equivalent to $\ln(r)$.

In effect, what happens in the calculation is that the cumulative probability $P(t)$ is calculated, i.e. the area under the probability curve up to a value t , and its value is then mapped on to a uniform random number distribution from which t is directly selected.

If $p(t)$ is the probability per unit time that an electron has a flight, of duration time t , terminated by some scattering process (i.e. has a drift time t in momentum space and its then scattered) then the solution of equation (III.1-2) in the term form

$$r = \int_0^1 p(t') dt' \quad (\text{III.1-6})$$

will enable a random distribution of such flight time to be generated.

Now suppose that the electron drifts for a time before being scattered and that this time consists of n tiny increments $\delta t_1, \delta t_2, \dots, \delta t_n$. The probability of the electron being scattered, within the time interval δt_i , is $\lambda(k) \delta t_i$, where $\lambda(k)$ is the total scattering rate defined by following equation:

$$\lambda(k) = \sum_{n=1}^N \lambda_n(k) \quad (\text{III.1-7})$$

where

$$\lambda_n(k) = \int S_n(k, k') dk \quad (\text{III.1-8})$$

the probability that there will not be any scattering during δt_i is thus $(1 - \lambda(k) \delta t_i)$. Therefore, the probability of an electron drifting in momentum space, for a time t , is

$$S(t) = \prod_{i=1}^n (1 - \lambda(k) \delta t_i) \quad (\text{III.1-9})$$

so that

$$\log\{S(t)\} = \sum_{i=1}^n \log(1 - \lambda(k) \delta t_i) \quad (\text{III.1-10})$$

However, since $\lambda(k) \delta t_i \ll 1$, equation (III.1-10) reduces to

$$\log\{S(t)\} = -\sum_{i=1}^n \delta t_i \lambda(k) \quad (\text{III.1-11})$$

which gives, immediately

$$S(t) = \exp \left\{ - \int_0^t \lambda(k) dt' \right\} \quad (\text{III.1-12})$$

where $\lambda(k)$ is a function of time through following equation:

$$k_r(t) = k_i - \frac{eF}{\hbar} t \quad (\text{III.1-13})$$

The probability density $p(t)$ is therefore

$$p(t) = \lambda(k) \exp \left\{ - \int_0^t \lambda(t) dt' \right\} \quad (\text{III.1-14})$$

Using equation(III.1-14) and a uniformly distribution random number distribution gives

$$r = 1 - \exp \left\{ - \int_0^t \lambda(k) dt' \right\} \quad (\text{III.1-15})$$

In the program, we include acoustic phonon scattering, polar optic phonon scattering, intervalley scattering, impurity scattering, alloy scattering and piezoelectric scattering.

III.2 MC results for some materials

In this section, the calculated scattering rates for some materials are shown that are used in the MC simulation. In Fig. III.2-1, the scattering rate due to impurity is shown as function of incident electron energy.

In Fig.III.2-2, we calculate the interface scattering rate as a function of rough length L in InGaAs/InAlAs system. The $\text{In}_{0.53}\text{Ga}_{0.47}\text{As}-\text{In}_{0.52}\text{Al}_{0.48}\text{As}$ modulation-doped (MD) heterostructures is important for FET applications. Since in an MD structure the electron motion occurs close to the interface, it is important to develop a model for the structural quality of an interface and use it to calculate the effects of interface roughness. Following Hong's work¹², the scattering rate due to interface roughness is given as:

$$W_{\text{inter}} = \frac{4\pi e^4 m^*}{\hbar^3 \epsilon_s^2} \left(N_{\text{acc}} + \frac{N_s}{2} \right)^2 \Delta^2 \int_0^\pi J_1(K_F L \sin \phi)^2 d\phi \quad (\text{III.2-1})$$

where J_1 is the first-order Bessel function, N_{acc} is the accumulation-layer charge density and N_s is the 2-D electron carrier density. In Fig.III-2, scattering rate for InGaAs-InAlAs interface is plotted as a function of L . It is important to realize that as L increases there is initially a sharp increase in W_{inter} (sharp decrease in mobility), but beyond $L \approx 50$ angstrom, W_{inter} become independent of L and

eventually begins to decrease (mobility start to increase). We can say that once the roughness increases beyond $1/K_F$, the electrons do not "feel" the surface roughness as much.

III.3 Velocity-field characteristics

In this section, simulated velocity-electric field curve for a few compound semiconductors are shown. The transport and other material parameters used in the simulation are shown in Table-1. The simulation results are obtained after 100,000 interaction. Some of the computation are carried out on the Air Force computer system.

Table 1. Parameters for III-V semiconductor.

| Parameter | GaAs | GaN | InAs | InP | GaSb | GaP |
|--|----------|----------|----------|----------|--------|----------|
| Density | 5.37 | 6.1 | 5.667 | 4.81 | 5.6137 | 5.477 |
| Velocity of Sound | 5.22 | 5 | 4.6 | 4.594 | | 5.847 |
| High freq. dielectric constant | 10.82 | 5.35 | 12.37 | 9.61 | 14.44 | 9.55 |
| Static dielectric constant | 12.53 | 9.5 | 15.15 | 12.56 | 15.69 | 12.4 |
| Polar optic phonon freq. | 5.37 | 15.1 | 4.54 | 8.79 | | 8.79 |
| Equivalent intervalley phonon freq. | 4.54 | 4.54 | 3.99 | 5.63 | | 5.63 |
| Non-equivalent intervalley phonon freq. | 4.54 | 4.54 | 3.99 | 5.63 | | 5.63 |
| Acoustic deformation potential | 7 | 12 | 5.8 | 3.4 | | 9 |
| Equivalent intervalley deformation potential | 1.00E+10 | 1.00E+10 | 1.00E+09 | 1.80E+09 | | 1.10E+09 |
| Non-equivalent intervalley deform. potential | 1.00E+10 | 1.00E+10 | 1.00E+09 | 9.00E+08 | | 6.00E+09 |
| Central valley effective mass | 0.067 | 0.13 | 0.0239 | 0.079 | 0.22 | 0.0925 |
| Satellite valley effective mass | 0.35 | 0.7 | 0.72 | 0.43 | 1.1 | 0.42 |
| Valley separation | 0.36 | 1.5 | 1.138 | 1.8 | | 1.466 |

| | | | | | | |
|------------------|------|------|--------|--------|---------|------|
| Band gap | 1.42 | 3.38 | 0.354 | 1.344 | 0.75 | 2.78 |
| Lattice constant | 5.56 | 4.5 | 6.0583 | 5.8687 | 6.09593 | 5.45 |

III.4 Impact Ionization Process

III.4.1 Some Previous Model

The impact-ionization (I.I) process is very important to the study of modern semiconductor devices, such as avalanche photo detectors, in which gain is provided by carrier multiplication. In an I.I process a charge carrier with high kinetic energy (“hot carrier”) collides with a second charge carrier, transferring its kinetic energy level. I.I process may be classified as band-band process or band trap process depending on whether the second carrier is initially in the valence band and makes a transition from the valence band to the conduction band, or whether it is initially at localized level (trap, donor, acceptor), and makes a transition to a band state. In HEMTs with narrow bandgap channel (like InAs, InGaAs), impact ionization can affect the device performance.

Wolff¹³ is perhaps the first to put forward a diffusion model for impact ionization with an argument that electrons gain energy gradually due to many collisions. Wolff’s model is likely to dominate at very high electric fields where the average electron energy is comparable to the ionization threshold energy. Shockley’s model¹⁴, often referred to as the lucky electron model, presumes that only those few electrons lucky enough to avoid collision gain sufficient energy from field for the ionization. These lucky electrons could be characterized by their ballistic movement. This ballistic movement is expected to be dominant at low electric field where the electron energy is very low compared to the ionization threshold energy.

The most widely used model, due to Baraff¹⁵, reduces to Shockley’s model in the low-field range while it converges to Wolff’s model in high-field limit. Recently, Ridley^{16,17} has extended Shockley’s lucky electron model by making a distinction between the rates of momentum and energy relaxation. This “lucky drift” model is intermediate between Shockley’s ballistic state and Wolff’s equilibrium state. All these model , however, assume that the mean free paths for scattering are constant and independent of energy. In general, the mean free paths do depend on energy. Furthermore, these models do not give the exact form of the energy distribution function.

Perhaps the first attempt for energy dependent results is that by Keldysh¹⁸ who solved the Boltzmann transport equation on the basis of an assumed form for the symmetric part of the energy distribution function. Chen and Tang have provided an explicit expression for energy-dependent mean free path that is required for the use of Keldysh's results. By using simple band structure and assume that collision of impact ionization do not dependent on the angle between initial and final state, Keldysh give a simple formula for the rate of impact ionization:

$$\frac{1}{\tau_{ii}(E)} = \begin{cases} 0, E \leq E_{th} \\ \frac{P}{\tau_{op}(E_{th})} \left[\frac{E - E_{th}}{E_{th}} \right]^2, E > E_{th} \end{cases} \quad (\text{III.4-1})$$

where E_{th} is a threshold energy and $1/\tau_{op}(E_{th})$ is the electron-optic-phonon scattering rate averaged over all electron wave vectors corresponding to the threshold energy E_{th} . Finally, P is a coefficient which we consider merely a fitting parameter, and it dependent with materials. Keldysh's model is often referred to as "hard threshold" model.

Another important model is "random-k" approximation by Kane¹⁹. In Kane's model, the energy-dependent rate for inelastic scattering of production of electron-hole pairs is computed by first-order perturbation theory using a screened Coulomb interaction with a frequency- and momentum- dependent dielectric function in the random-phase approximation. The threshold for momentum-conserving pair creation is found to very close to that determined by energy conversion alone. Using this "random-k" method, the scattering rate for primary holes is obtained and found to be almost identical with that for primary electrons of comparable energy. The secondary-particle energy distribution function are also determined for primary holes and electrons. One-electron state-density structure is prominent in these distributions.

III.4.2 Present Work

Our work is partly based on the models proposed by Kane and Kunikiyo²⁰. First of all, we just interest at the electron behaves near the threshold energy, so we assume that the initial state of secondary electron is near the top of valence band, and we do not care about the hole's energy. For simplify the calculation, we begin with simple band structure.

The impact ionization rate is calculated by the use of Fermi's golden rule. In the impact-ionization process, an electron in the conduction band interacts with an electron in the valence band via the screened Coulomb potential $V(r-r')$:

$$V(r-r') = \frac{e^2}{4\pi\epsilon(q, \omega)} \frac{1}{|r-r'|} \quad (\text{III.4-2})$$

If state 1 (a Bloch state with band index γ_1 and wave vector k_1) and 2 are the states of the initial electron in the conduction band and valence band before the transition, respectively, and 1' and 2' are those of the final electrons in the conduction band after the transition, the ionization rate $1/\tau_{\text{II}}(1,2 \rightarrow 1',2')$ is expressed by:

$$\frac{1}{\tau_{\text{II}}(1,2 \rightarrow 1',2')} = \frac{2\pi}{\hbar} \left[|M_a|^2 + |M_b|^2 + |M_a - M_b|^2 \right] * \delta(E_1 + E_2 - E_{1'} - E_{2'}) \quad (\text{III.4-3})$$

where $\delta(E)$ is the energy conserving delta function and $E_i (i=1,2,1',2')$ denote the energy of each electron. The direct matrix element M_a and the exchange matrix element M_b are given by:

$$M_a = \left\langle \psi_1(r_1) \psi_2(r_2) \left| \frac{e^2}{4\pi\epsilon(q, \omega) |r_1 - r_2|} \right| \psi_1(r_1) \psi_2(r_2) \right\rangle \quad (\text{III.4-3.1})$$

$$M_b = \left\langle \psi_2(r_1) \psi_1(r_2) \left| \frac{e^2}{4\pi\epsilon(q, \omega) |r_1 - r_2|} \right| \psi_1(r_1) \psi_2(r_2) \right\rangle \quad (\text{III.4-3.2})$$

where $\epsilon(q, \omega)$ is a dielectric function depending on both wave vector and transition energy. $|M_a|^2 + |M_b|^2$ in equation (III.4-3) denotes the matrix element in the case that the spin of electron 1 and 2 is different, while $|M_a - M_b|^2$ denotes the matrix element in the case that spin of electron 1 and 2 is the same. The wave function of electron $\psi_{v,k}$ is expressed by the Fourier series over reciprocal lattice vector G

$$\psi_{v,k}(r) = \frac{1}{\sqrt{V_0}} \sum_G A_{v,k}(G) e^{i(k+G)r} \quad (\text{III.4-4})$$

where k_v is the wave vector associated with energy band v . The wave function are obtained from the empirical local pseudopotential method. If we represent $1/|r_1 - r_2|$ by the Fourier expansion, then

$$\frac{1}{|r_1 - r_2|} = \frac{1}{V_0} \sum_q \frac{4\pi}{q^2} e^{iq \cdot (r_1 - r_2)} \quad (\text{III.4-5})$$

By substituting (III.4-4) and (III.4-5) into equation (III.4-3) and (III.4-3.1), (III.4-3.2), we get:

$$M_a = \frac{e^2}{\varepsilon(q, \omega)} \frac{1}{V_0} \sum_{G_1, G_2, G_1', G_2'} A_{k_1, \gamma_1}^*(G_1) A_{k_2, \gamma_2}^*(G_2) \frac{1}{|k_1' + G_1' - k_1 - G_1|^2} \quad (III.4-6)$$

$$* \delta(-k_1' - G_1' + k_1 + G_1 - k_2' - G_2' + k_2 + G_2)$$

$$M_b = \frac{e^2}{\varepsilon(q, \omega)} \frac{1}{V_0} \sum_{G_1, G_2, G_1', G_2'} A_{k_2, \gamma_2}^*(G_1) A_{k_1, \gamma_1}^*(G_2) \frac{1}{|k_1' + G_1' - k_1 - G_1|^2} \quad (III.4-7)$$

$$* \delta(-k_1' - G_1' + k_1 + G_1 - k_2' - G_2' + k_2 + G_2)$$

where δ is the wave vector conserving delta function. Probability of transition from initial state 1 is obtained by equation (III.4-3) summing over k_2, k_1', k_2' :

$$W_{11}(1) = \sum_{k_2, k_1', k_2'} W_{11}(1, 2 \rightarrow 1', 2') \quad (III.4-8)$$

The rate is calculated numerically in such a way as to conserve both momentum and energy of carrier in each transition.

Using the two-band approximation²¹, we calculate the matrix element M_a and M_b , unlike assumption that the matrix element is constant so it can be brought out of the sum motion in calculation of scattering rate, which are dependent not only on the initial and final energy, but also on the angle between the wave vector of those two states. The matrix elements can be written as:

$$|M_a|^2 = \frac{1}{4} \left(\frac{e^2}{\varepsilon} \frac{1}{V_0} \right)^2 \left\{ (a_2 + b_2 \cos \theta_2 + c_2 \cos^2 \theta_2) * (a_1 + b_1 \cos \theta_1 + c_1 \cos^2 \theta_1) \right\} \left| \frac{1}{(k_1' + G_1' - k_1 - G_1)^2} \right|^2 \quad (III.4-9)$$

$$|M_b|^2 = \frac{1}{4} \left(\frac{e^2}{\varepsilon} \frac{1}{V_0} \right)^2 \left\{ (a_2 + b_2 \cos \theta_2 + c_2 \cos^2 \theta_2) * (a_1 + b_1 \cos \theta_1 + c_1 \cos^2 \theta_1) \right\} \left| \frac{1}{(k_1' + G_1' - k_1 - G_1)^2} \right|^2 \quad (III.4-10)$$

where θ_i ($i=1, 2, 1', 2'$) is the angle between different states 1, 2, 1' and 2'. The a, b and c are constants which dependent on the state energy and the effective mass of different valley in conduction band. Substituting (III.4-9) and (III.4-10) into equation (III.4-3) and (III.4-8), and change the sum in (III.4-8) to integral, we obtain:

$$W(k_1, k_2, k_1', k_2') = \frac{2\pi}{\hbar} |M|^2 \delta(E_1 + E_2 - E_1' - E_2') \quad (\text{III.4-11})$$

$$W(k_1) = \frac{2\pi}{\hbar} \left(\frac{V_0}{(2\pi)^3} \right)^2 \iiint_{\Omega} dk_2 dk_1' dk_2' |M|^2 \delta(E_1 + E_2 - E_1' - E_2') \quad (\text{III.4-12})$$

Following the usual definition, the impact ionization coefficient α is given as the ratio of the average probability of impact ionization to the average of electron drift velocity:

$$\alpha = \int_0^{\infty} W_{ii}(\varepsilon) \Omega(\varepsilon) f_0(\varepsilon) d\varepsilon / v_d \int_0^{\infty} \Omega(\varepsilon) f_0(\varepsilon) d\varepsilon \quad (\text{III.4-13})$$

where $W_{ii}(\varepsilon)$ is the impact ionization scattering rate, $\Omega(\varepsilon)$ is the density of states, and the v_d is the electron drift velocity. For the range of electric field values relevant to the impact ionization process, v_d can be approximated by the saturation velocity v_s .

III.4.3 Results

In this section, the impact ionization rate calculated from the formal principle is shown in Figs III.4.1-5 for a few compound semiconductor.

From the results for impact-ionization rate calculation, we observe that the scattering rate is strongly dependent on the band gap of different materials. When the band gap of material is around 1.0 eV, the scattering rate increases with the of the initial energy. However, it is quite different in the narrow-bandgap semiconductors, such as InAs and InSb. For these materials the scattering rate increase rapidly near the threshold energy, and then decreases.

In Fig. III.4-6 the velocity-electric field characteristics for GaAs is plotted with temperature as a parameter. Both with and without impact ionization results are shown. In Fig. III.4-7 and Fig. III.4-8, the the velocity-electric field characteristics for $\text{In}_{0.53}\text{Ga}_{0.47}\text{As}$ and InAs are shown. For wide bandgap materials (like GaAs), impact ionization does not modify the velocity-electric field characteristics. However, with decreasing band gap, the contribution of impact ionization in the velocity-electric field characteristics become obvious as is shown in In Fig. III.4-7 and Fig. III.4-8. In narrow-bandgap material, the velocity-electric field characteristics can be divided electric field in the first regions, under low electric field, the electron energy is lower than impact ionization threshold implying that impact ionization is not occurring and will result in drift velocity comparable for without impact ionization. In the second region, under high electric field, impact-ionization scattering

becomes dominant scattering mechanism, and the drift velocity in the process of impact ionization is lower than without I.I. calculation in this case, electrons loose energy during the impact ionization process, and transfer energy and momentum to the secondary electron-hole pair. In the third region, under very high electric field, some electrons gain high energy from electric field during free flight and does not "feel" the affect of electrons in valence band, therefore impact ionization does not show affect the velocity-electric field characteristics. Moreover, with increasing temperature, the incident electron energy increase making impact ionization more probable as is shown in high temperature plots in Fig.III.4-6.

In Fig.III.4-9, the impact ionization coefficients is plotted as a function of electric field at different temperature for $\text{In}_{0.53}\text{Ga}_{0.47}\text{As}$. The effect of temperature is similar as that of Fig.III.4-6. In Fig.III.4-10, Fig.III.4-11 and Fig.III.4-12, low field mobility, peak velocity and saturation velocity are plotted as a function of In mole fraction for $\text{In}_x\text{Ga}_{1-x}\text{As}$ system at room temperature. In the inside, the difference between the results with and without impact ionization results are shown. As observed, impact ionization decreases μ , v_p and v_s with increasing In mole fraction. Beyond the In fraction of 80%, the difference is quiet noticeable and will affect device performance.²²⁻³⁰

IV Device characterization

IV.1 DC Analysis

The current-voltage characteristic is evaluated in the reduced potential scheme³¹. To facilitate calculation the self-consistently calculated x_{av} and E_F are expressed in the following functional forms:

$$x_{av}(n_s(x)) = a + b \ln(n_s(x)) \quad (\text{\AA}) \quad (\text{IV.1-1})$$

$$\begin{aligned} E_F(n_s(x)) &= E_{FO} + \gamma \ln(n_s(x)) \\ &= E_{FO} + \gamma \Delta E_F(n_s(x)) \quad (\text{eV}) \end{aligned} \quad (\text{IV.1-2})$$

where x represents the position in the channel with respect to the source at $x=0$. The numerical values of a , b , E_{FO} and γ are obtained by fitting the x_{av} vs n_s and E_F vs n_s curves obtained by solving Schroedinger and Poisson's equations. The 2DEG concentration using reduced potentials can be written as

$$n_s(x=0) = \frac{\varepsilon \cdot [V_{gs} - \Delta E_F(x=0)/q - V_T]}{q \cdot d_{eff}} = \frac{\varepsilon \cdot |V_T| \cdot s}{q \cdot d_{eff}} \quad (\text{IV.1-3})$$

where the reduced potential at the source end of the device is:

$$s = \frac{V_{gs} - \Delta E_F(x=0)/q - V_T}{|V_T|} \quad (\text{IV.1-4})$$

$d_{eff} = d - \Delta d(n_s(x=0))$ and d is the thickness of *AlGaAsSb* layer, $\Delta d(n_s(x=0))$ denotes the effective channel width at the source and equals $(\epsilon_{AlGaAsSb} / \epsilon_{InGaAs}) \cdot x_{av}(n_s(x=0))$. It should be mentioned that $\Delta d(n_s(x))$ is a function of n_s and is properly updated depending on $n_s(x)$ (the channel is narrower at the source ($x=0$) and widens as one move towards the drain region ($x=L_g$)). Using the definition of s , the gate-source voltage may be written as

$$V_{gs} = V_T + s \cdot |V_T| + \gamma \cdot \ln\left(\frac{\epsilon \cdot s \cdot |V_T|}{q \cdot d_{eff}}\right) \quad (\text{IV.1-5})$$

where the threshold voltage $V_T = \phi_b - q/\epsilon(N_d d_d^2/2 + N_d d_d d_i + N_d d^2/2)$. ϕ_b , d_d , d_i and N_d are the Schottky barrier height, the thickness of *AlGaAsSb* donor layer, the thickness of spacer layer and doping concentration of *AlGaAsSb* layer, respectively.

$$V_{gs} = V_T + s \cdot |V_T| + \gamma \ln(n_s(x=0)) \quad (\text{IV.1-6})$$

The saturation drain current $I_{c,sat}$ in the reduced potential scheme can be written as:

$$I_{c,sat} = \frac{G_0 |V_T|}{\epsilon_0 L_1} \left(\frac{s}{2} \sqrt{s^2 - p^2} - \frac{p^2}{2} \ln\left(\frac{s + \sqrt{s^2 - p^2}}{p}\right) + \frac{\gamma}{|V_T|} [\sqrt{s^2 - p^2} - p \cdot \cos^{-1}(\frac{p}{s})] \right) \quad (\text{IV.1-7})$$

where $p = \frac{V_{gs} - V(x=L_1) - \Delta E_F(x=L_1)/q - V_T}{|V_T|}$. $v_d = \frac{v_s \cdot \epsilon}{\sqrt{(\frac{v_s}{\mu_0})^2 + \epsilon^2}} = \frac{v_s \cdot \epsilon}{\sqrt{\epsilon_0^2 + \epsilon^2}}$ is the assumed

velocity-electric field characteristic. The magnitude of the drain-source voltage for a given drain current is given by

$$V_{ds} = (V_g - V_T - \frac{\Delta E_F(x=0)}{q}) + I_c \cdot (R_d + R_s) - \frac{I_c}{G_0} \cdot [1 + \{(\frac{G_0 |V_T|}{I_c} - 1)^\alpha - \frac{v_s L_g G_0}{\mu_0 I_c}\}^\beta] \quad (\text{IV.1-8})$$

for $0 < I_c < I_{c,sat}$. When the device is in saturation, $I_c > I_{c,sat}$, the drain-source voltage can be written as

$$V_{ds} = (V_g - V_T - \frac{\Delta E_F(x=L_g)}{q}) + I_c \cdot (1/G_0 - R_d - R_s) + \frac{2 \cdot d_{eff} \cdot \epsilon_0}{\pi} \sinh\left(\frac{\pi}{2 \cdot d_{eff}} \cdot [L_g - \frac{I_c}{\epsilon_0 G_0} (\frac{s G_0 |V_T|}{I_c} - 1)^\alpha]\right)$$

Here I_c is the channel current, $\beta = 2/3$, $G_0 = (\epsilon \cdot Z \cdot v_s / d_{eff})$, L_g and Z are the gate length and width, respectively. R_s is the source resistance and R_d is the drain resistance. To fit the experimental I-V curves in the saturation region, a parasitic resistance R_p is introduced. The modified drain-source

current $I_{ds} = I_c + V_{ds}/R_p$. The evaluation of the dc small signal parameters, namely (a) transconductance g_m , (b) drain resistance r_d and (c) the gate capacitance C_{GS} are evaluated by using the method detailed in Ref.³².

IV.2 Results and discussion

An $\text{In}_{0.53}\text{Ga}_{0.47}\text{As}/\text{Al}_{0.6}\text{Ga}_{0.4}\text{AsSb}$ double barrier HEMT with a QW width of 100\AA is considered. The calculated band offset is 0.984eV . A low field mobility of $53,000\text{ cm}^2/\text{V-s}$ and a saturation velocity of $3.2 \times 10^7\text{ cm/s}$ is used in the simulation⁷. The donor doping density of $2 \times 10^{18}\text{ cm}^{-3}$ with a doped epilayer thickness of 350\AA is used.

The current-voltage characteristics are shown in Fig. IV.1-1 where the gate bias is varied from -1.0 V to 0V at an interval of 0.5V . An impact ionization constant of $5 \times 10^5/\text{cm}$ is used in the computation. The I-V curves for higher gate bias clearly shows the effect of impact ionization.

Fig. IV.1-2, shows the calculated transconductance as a function of the gate bias. A maximum transconductance of 800 mS/mA , for a gate length of $1\mu\text{m}$, is obtained for the unoptimized structure.

IV.3 Small signal analysis

HEMTs using InGaAs as the channel material are ideal for their superior transport properties and are being used in millimeter wave technology. However, the narrow band gap of InGaAs makes it susceptible to impact ionization. Impact ionization in HEMTs gives rise to kinks in the current-voltage (I-V) characteristics, current instability and hysteresis along with high gate current. Impact ionization is also known to increase device noise³³. Impact ionization is initiated at the drain end of the gate where a large electric field exists. Kruppa et. al.³⁴ have investigated impact ionization in AlSb/InAs HEMTs and have modeled the inductive behavior of S_{22} by a series RL circuit at the drain end of the small signal equivalent circuit. Reuter et. al. suggested a voltage controlled current source in an RC environment to explain the inductive S_{22} due to impact ionization. The current source is controlled by the voltage across the gate-drain region namely the voltage across C_{GD} . It should be mention that the inductive nature of S_{22} is observed only at low frequencies suggesting that impact ionization may be dominant only at the low frequency end of the spectrum (upto a few GHz). This frequency dependence of the impact ionization current source is obtained by the arrangement of the RC network in an otherwise frequency independent g_m . Though the inductive nature of S_{22} indicates

the presence of impact ionization an exact quantification is extremely difficult. Reuter et. al.³⁴ used a genetic optimization technique to obtain the impact ionization induced transconductance g_m^i . In this paper, we report a technique that enables one to determine g_m^i directly from experimental data and does not require optimization³⁴.

The lattice matched InP based HEMT structure is described in Ref.³⁵. The gate length and widths are 0.15 μm and 40 μm respectively. The measurements and simulations are carried out at room temperature. The small signal equivalent circuit reported by Webster et. al.³⁶ is modified by incorporating the impact ionization current source at the output with the RC branch as shown in Fig.IV.2-1. The optimized small signal equivalent circuit is shown for bias conditions $V_G=-0.3\text{V}$, $I_G=-6.4\text{ }\mu\text{A}$, $V_D=1.5\text{V}$ and $I_D=7.8\text{mA}$. The circuit is optimized to the experimental S-parameters over the frequency range of 0.04GHz to 40 GHz. In Fig.IV.2-2, the experimental and modeled S-parameters are shown and the agreement over the frequency range is excellent. The inductive S_{22} carries the signature of impact ionization as modeled by the voltage dependent current source at the output in accordance with the work reported by Reuter et.al.³⁴. Though an optimization of the circuit may fit the measured S-parameters, does not provide an experimental measure of g_m^i .

In Fig.IV.2-3, the experimental Y_{21} is shown as a function of frequency, for $V_D = 1.5\text{V}$ with varying gate potential. An interesting point to note is the sudden increase in Y_{21} at $\omega \rightarrow 0$ or near dc. This sudden variation can be explained by inspecting Y_{21} at dc. For the intrinsic circuit, $Y_{21}(\omega)$ is expressed as:

$$Y_{21} = g_m + g_m^i \left(1 - \frac{j\omega C_{im}}{\frac{1}{R_{im}} + j\omega C_{im}} - j\omega C_{GS} - \frac{(g_m + j\omega C_{DC}) \left(\frac{1}{R_{GS}} + j\omega C_{GS} \right)}{j\omega (C_{DC} + C_{GS}) + \frac{1}{R_i} + \frac{1}{R_{GS}}} \right) \quad (\text{IV.2-1})$$

where ω is the angular frequency and the rest of the parameters are defined in the small signal equivalent circuit shown in Fig.IV.2-1. Under dc condition and with $R_{ds} \gg R_i$ the above expression reduces to:

$$Y_{21}(\omega \rightarrow 0) = g_m + g_m^i. \quad (\text{IV.2-2})$$

The expression strongly suggests that the sudden change in $Y_{21}(\omega)$ as ω approaches 0 is due to g_m^i . This provides a means to determine g_m^i purely from experimental data for the first time. In Fig.IV.2-4, the extracted g_m^i is shown as a function of gate and drain bias. For a given gate bias g_m^i increases

monotonically with drain bias and may be explained by accounting for the increasing gate-drain electric field. The variation of g_m^i with gate bias for a given drain potential peaks at a certain gate bias. This phenomena is due to the dependence of g_m^i upon both the gate-drain electric field and the drain current. For a given drain bias the gate-drain electric field E_{DG} increases as the gate bias becomes more negative. Consequently the ionization constant increases. However, with more negative gate bias the drain current decreases and eventually at some negative gate bias the channel is pinched off. Therefore, at some intermediate gate voltage g_m^i will maximize and in Fig.IV.2-4(b), its dependence is shown for different drain bias.

V Noise

Modeling of noise is based upon a self-consistent solution of Schroedinger and Poisson's equations. The analysis of noise is based on the identification of the different noise source that present in the channel,^{9,12} namely (a) Johnson Noise in the ohmic region, (b) noise associated with spontaneous generation of the dipole layers in the saturation region, (c) gate noise due to elementary voltage fluctuations in the channel and (d) induced gate noise in the saturation region. Moreover, noise due to impact ionization is introduced through g_m^i . The noise spectrum of the impact ionization current source is assumed to be white (in Sec. 6 this aspect of the study is discussed).

The self-consistent noise model³⁷ is extended to calculate noise properties in InP-based HEMTs. By accounting for the noise sources and matching the optimized external source impedance to the transistor, the minimum noise figure F_{min} and minimum noise temperature T_{min} are calculated as³⁷ :

$$F_{min} = 1 + 2 \cdot g_n (R_c + \sqrt{R_c^2 + \frac{r_n}{g_n}}) \quad (V-1.1)$$

$$T_{min} = 2 \cdot T \cdot g_n \cdot (R_c + R_{c,opt}) \quad (K) \quad (V-1.2)$$

where

$$g_n = g_m \cdot \left(\frac{f}{f_T}\right)^2 \cdot (R + P - 2C\sqrt{PR}) \quad (V-1.3)$$

$$R_c = R_s + R_g + R_i \quad (V-1.4)$$

where r_n is the noise resistance, R_c is the correlation resistance, R_s and R_g are the source and drain resistances and $R_i = L_g/v_s$, C_{gs} is the gate charging resistance and L_g represents the length of the gate. P , R and C represent the noise coefficients and $R_{c,opt}$ is the optimal external source resistance.

g_m and C_{gs} are the device transconductance and gate capacitance, respectively. T is the operating temperature in 0 K.

V.1 Results and Discussion

An $\text{In}_{0.53}\text{Ga}_{0.47}\text{As}/\text{Al}_{0.6}\text{Ga}_{0.4}\text{AsSb}$ double barrier HEMT with a QW width of 100\AA is considered. The calculated band offset is 0.984eV . A low field mobility of $53,000\text{ cm}^2/\text{V-s}$ and a saturation velocity of $3.2 \times 10^7\text{ cm/s}$ is used in the simulation³⁸. The donor doping density of $2 \times 10^{18}\text{ cm}^{-3}$ with a doped epilayer thickness of 350\AA is used.

Fig.V-1, shows the minimum noise figure as a function of frequency. The gate and drain biases are 1.5V and 2.0V , respectively. As observed, the minimum noise figure is less than 1dB at 60GHz . An optimization based upon (a) QW width (b) gate length and (c) doped epilayer thickness will enable the realization of lower F_{\min} upto 100GHz . Moreover, the role of gate drain separation on noise due to impact ionization is under investigation.

VI. Frequency Spectrum of Impact Ionization

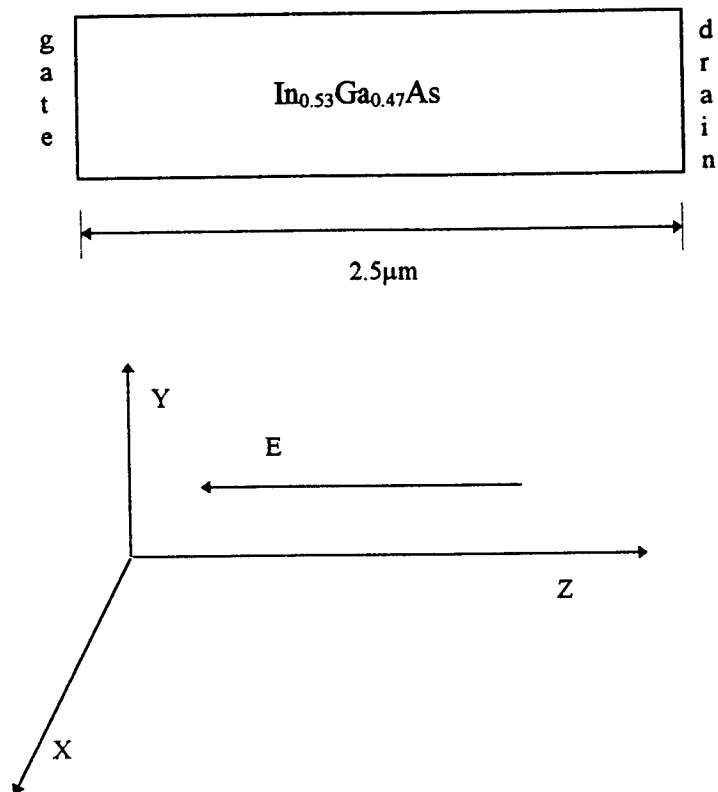
In order to investigate the frequency dependence of the impact ionization process two different studies are carried out, namely (a) time dependent solution of Boltzman transport equation using Monte Carlo simulation and (b) analytical study using continuity equation.

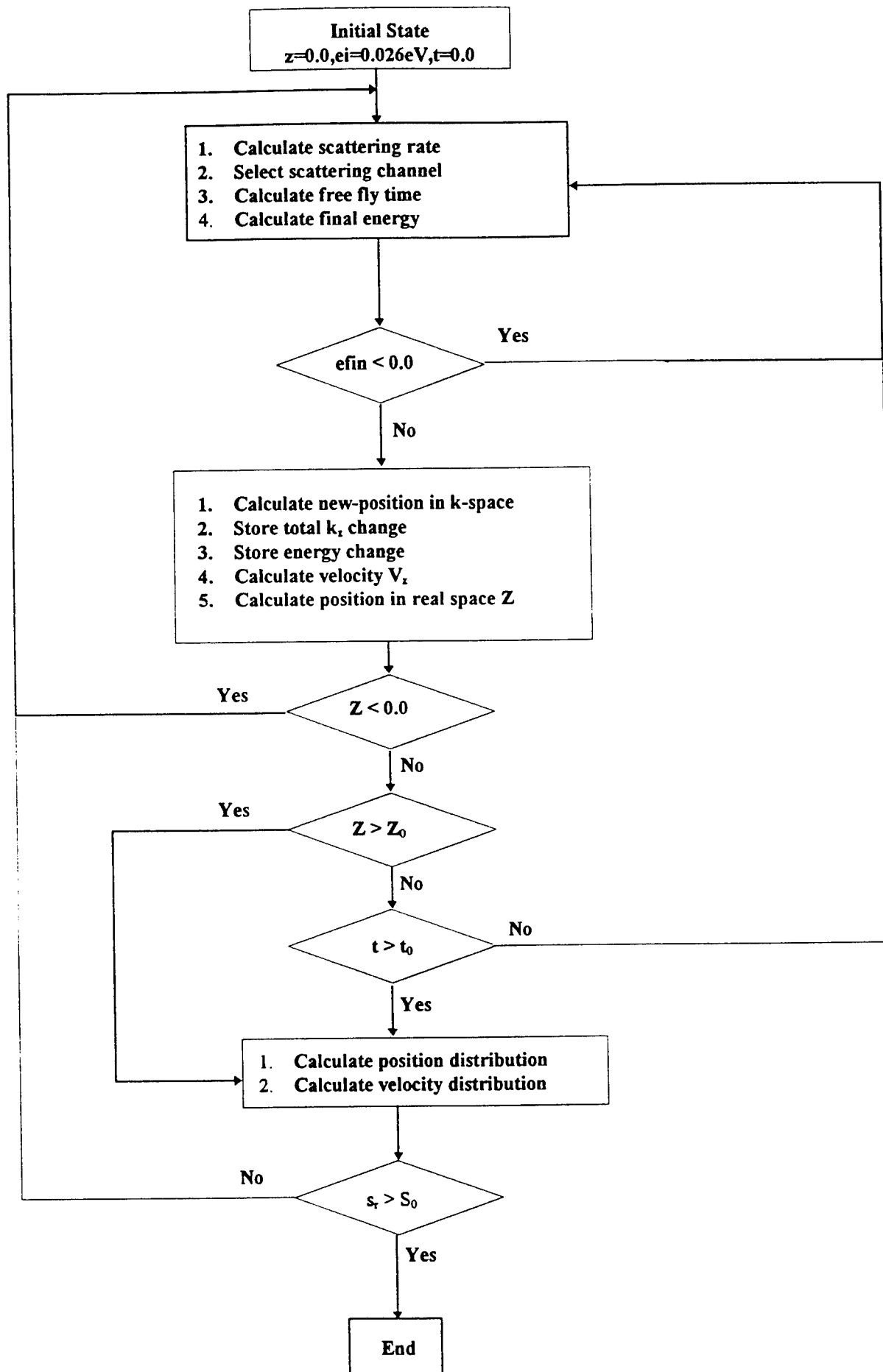
VI.1 MC Simulation of Time Dependent Electrons Distribution

In this section, some results of MC simulation for time dependent electrons distribution in k -space as well as in real space are shown. The electric field is applied along the z direction. The region between gate and drain in HEMTs where the reverse electric field is extremely high and impact ionization could happen is under investigation.

In Fig.VI.1-1, electron position in k -space is plotted with time period as a parameter. In Fig.sVI.1.2-3 the distribution of electron position in real space is plotted as a function of time period as well as the distribution of drift velocity. With increasing time, the electrons distribution in k -space and in real space become broadened. The average velocity of electrons decrease with increasing time. With longer time period (correspond to low frequency), the electrons have to suffer more scattering events, thus their energy (vector in k -space) and their position in real space become more

random and consequently velocity is reduced. This may result in an additional phase shift in the drain current with the magnitude of the current remaining the same.





VL2 Frequency Dependent Drain Current Due to Impact Ionization

Impact ionization process is important in narrow bandgap channel HEMTs and needs to be investigated.

Following the approach of H.K. Gummel and J.L.Blue⁴⁰, we calculate the hole current (which equal to electron current due to impact ionization) as a function of drain current as well as frequency. Equal ionization coefficients were assumed for holes and electrons and sum of the hole and electron currents in the region between gate and drain was assumed constant.

From the continuity equations for holes and electrons, we obtain:

$$\frac{\partial}{\partial t}(p + n) + \frac{\partial}{\partial z}(v_h p - v_e n) = 2(g + \gamma) \quad (\text{VI.2-1})$$

where v_h , v_e is velocity of holes and electrons respect, the noise-less carrier generation rate $g = \alpha v_h p + \beta v_e n$, and γ is a stochastic generation rate (which we assume it is equal to zero).

By using saturation velocity approximation which is reasonable under the high electric field can make impact ionization happen and assume p is negligible, we get:

$$I_{\text{tot}} = qv_e n + \frac{\partial}{\partial t} \epsilon E \quad (\text{VI.2-2})$$

After some steps, we obtain the drain current is:

$$I(x, t) = I_0 e^{\frac{1}{v_e}(j\omega - 2\beta v_e)(Z - Z_0) + j\omega t} \quad (\text{VI.2-3})$$

where I_0 is the current inject from gate end which we assume it is constant. So the hole current in the gate-drain region due to carrier multiplication can be write as:

$$I_p(t) = \frac{1}{L_{gd}} \int_0^{L_{gd}} I_0 e^{\frac{1}{v_e}(j\omega - 2\beta v_e)Z + j\omega t} dz - I_0 \quad (\text{VI.2-4})$$

So we can obtain frequency dependent hole current as:

$$I_p(\omega) = \frac{I_0}{L_{gd}(4\beta^2 + \frac{\omega^2}{v_e^2})} (\cos \omega t + j \sin \omega t)(A + jB) \quad (\text{VI.2-5})$$

where

$$A = 2\beta(1 - \cos \frac{\omega L_{gd}}{v_e} e^{-2\beta L_{gd}}) + \frac{\omega}{v_e} e^{-2\beta L_{gd}} \sin \frac{\omega L_{gd}}{v_e} - 1 \quad (\text{VI.2-5.1})$$

$$B = \frac{\omega L_{gd}}{v_e} \left(1 - \cos \frac{\omega L_{gd}}{v_e} e^{-2\beta L_{gd}}\right) - 2\beta e^{-2\beta L_{gd}} \sin \frac{\omega L_{gd}}{v_e} \quad (\text{VI.2-5.2})$$

From Figs. 4-1 and 4-2, we see that there is a significant change of the current magnitude above v_e (around 100GHz), implying that I.I should be taken into account above this frequency.

VII Quaternary channels

$In_xGa_{1-x}As_{1-y}Sb_y$ lattice matched to $Al_aGa_{1-a}As_{1-b}Sb_b$ is investigated as a possible candidate for channel material for high performance HEMTs. In the following few paragraphs a justification for the proposed quaternary channel material is provided.

For an In mole fraction of 0.3 and a lattice constant of 5.948 \AA (lattice matched to $AlAsSb$) ΔE_c equals 0.9eV. The corresponding Sb mole fraction is $(y = (L_{xy} - 5.6533 - 0.4051 x)/(0.4426 - 0.0216 x))$, where L_{xy} is the lattice constant of the quaternary) is 0.4 yielding a band gap of 0.6 eV. The corresponding Γ -L separation is a modest 0.6 eV. It should be noted that the maximum ΔE_c for lattice matched $InAs/AlAsSb$ is 1.1 eV with the $InAs$ band gap of 0.36 eV. Therefore, the band gap improves from 0.36eV to 0.6 eV without sacrificing ΔE_c if $InAs$ is substituted by $In_xGa_{1-x}As_{1-y}Sb_y$. A deep QW guarantees electron confinement along with a higher 2DEG concentration, however, a decrease of ΔE_c from 1.1 eV to 0.9 eV does not radically change the 2DEG concentration as the higher sub-bands are sparsely populated. On the other hand, an increase of the band gap from 0.36eV to 0.6 eV will require that the electron gain an additional 0.36 eV ($=1.5*(0.6-0.36)$) of energy, from the applied field, making possible a great reduction in impact ionization.

The room temperature low field mobility $\mu(x,y)$ of $In_xGa_{1-x}As_{1-y}Sb_y$ is obtained by using the following relationship:

$$\mu(x,y) = 8.5 \times 10^3 + 195 \times 10^4 x - 3.5 \times 10^3 y + 5.55 \times 10^4 xy \quad (\text{VII-1})$$

The relationship is obtained by using the interpolation formulation reported by Moon et. al. [9]. Proper study of low field mobility should be carried out by using a Monte Carlo simulation. However, the above equation provides an approximate number so that the material under

investigation can at least be compared to $In_xGa_{1-x}As$. Eqn. 3, for $In_{0.53}Ga_{0.47}As$ provides a low field mobility value of $18 \times 10^4 \text{ cm}^2/V\text{-s}$ in accordance with the experimental value of $14 \times 10^4 \text{ cm}^2/V\text{-s}$ whereas μ or $In_{0.3}Ga_{0.7}As_{0.4}Sb_{0.6}$ equals $2.22 \times 10^4 \text{ cm}^2/V\text{-s}$, better than the low field mobility of $In_{0.53}Ga_{0.47}As$, however, much less than that of $InAs$.

$In_xGa_{1-x}As_ySb_{1-y}$ seems to be a good candidate for channel material in HEMTs since the band gap and ΔE_c , of the quaternary channel material, is large enough to suppress impact ionization and provide a high 2DEG concentration with good confinement. Also, the mobility is higher than that of $In_{0.53}Ga_{0.47}As$. Providing an f_T of 500 GHz for a $0.1 \mu m$ gate.

In order to justify the use of the above empirical relationship for low field mobility the velocity-electric field for quaternaries are determined. Given the alloy scattering rate of the binaries the alloy scattering rate $P(x,y)$ of the quaternary $A_xB_{1-x}C_yD_{1-y}$, is suggested to be:

$$P(x, y) = (1 - x)yP_{BC} + (1 - x)(1 - y)P_{BD} + xyP_{AC} + x(1 - y)P_{AD} \quad (\text{VII-2})$$

where x and y are the mole fractions in $In_xGa_{1-x}As_ySb_{1-y}$ and P_{AB} are the alloy scattering rate of the binaries. In Fig. VII.1, the alloy scattering rate in $In_xGa_{1-x}As_ySb_{1-y}$ as a function of x and y are shown.

In Fig. VII-2, the drift velocity is plotted as a function of electric field for different channel materials. InGaAsSb system has higher low field mobility and saturation velocity than InGaAs system. The low field mobility extracted from MC calculation agrees quite well with the data obtained from the empirical formulation.

The use of a quaternary channel can be made more attractive if the following scheme is used. In alloys of InGaAsSb alloy scattering will be one of the dominant process. If thin multiple layers of InAs-GaSb are grown till the same channel width is obtained it may provide a better velocity-electric field characteristic. In the proposed growth, alloy scattering is replaced by interface scattering. However, as there are less interfaces channel transport will improve. In Fig. VII-3, the velocity-electric field characteristics for alloyed and superlattice (SL) channel are compared. The SL channel has improved low field mobility and peak velocity.

VIII. Conclusion

A complete model that includes (a) quantum well calculations (b) transport and (c) noise for deep quantum well HEMT is presented. Time independent Boltzman transport equation (BTE) is solved using ensemble Monte Carlo to study the velocity-electric field characteristic in $\text{In}_x\text{Ga}_{1-x}\text{As}$. Impact ionization in narrow band gap channel gives rise to a degradation in the low field mobility and peak velocity. The solution of the time dependent BTE indicates that impact ionization process is frequency dependent and this dependence is also a function of drain current and source drain spacing. $\text{In}_x\text{Ga}_{1-x}\text{As}_y\text{Sb}_{1-y}$ provides excellent transport properties and may be used as a channel material in HEMTs. The usefulness of the quaternary channel can be further enhanced if a superlattice channel is used instead of an alloyed channel.

IX. Figures

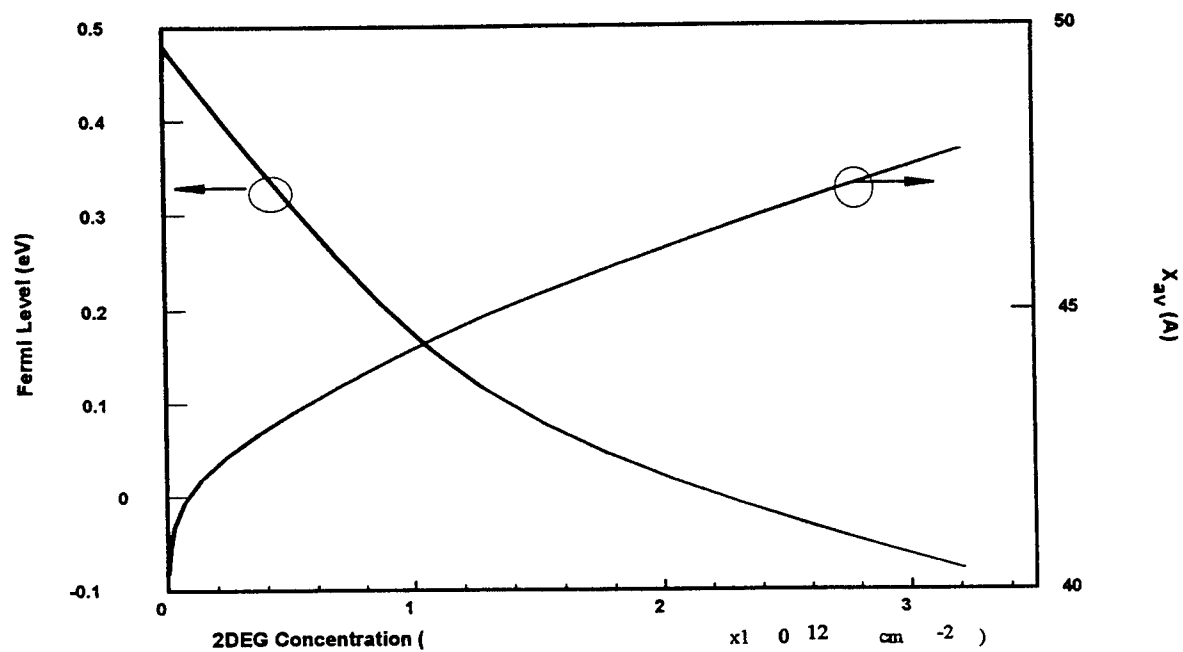


Fig.II-1 Fermi level and X_{av} as a function of 2DEG concentration at $t=300\text{K}$

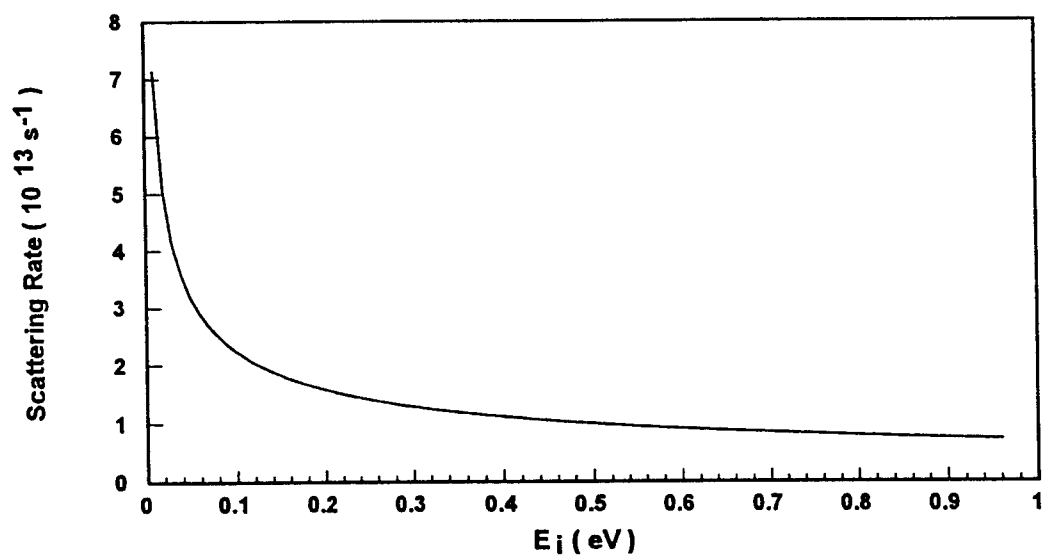


Fig.III.2-1 Impurity scattering rate calculation for GaAs

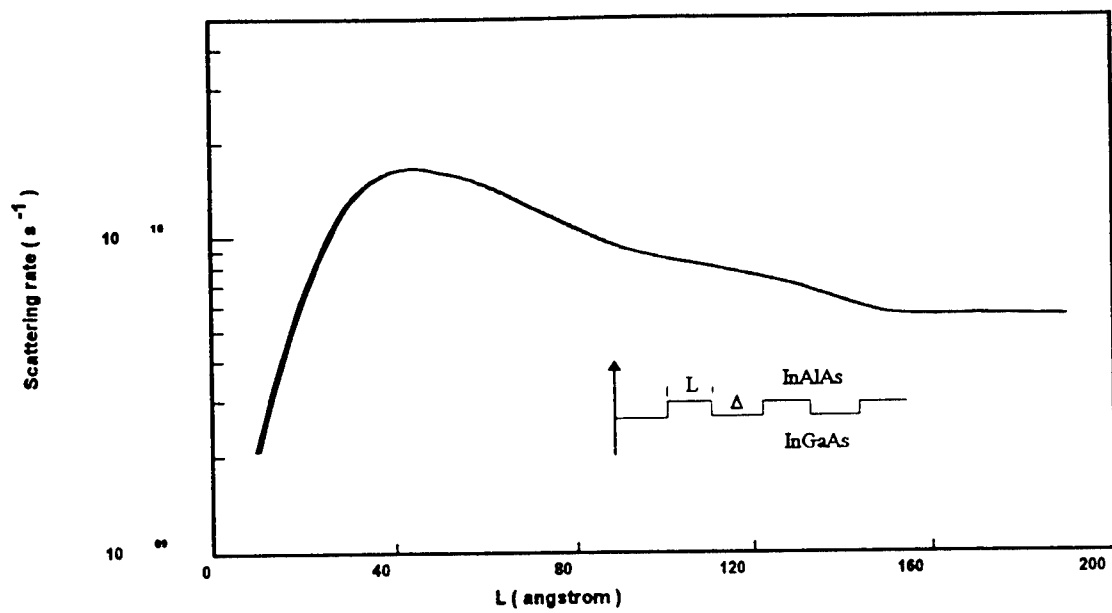


Fig. III.2-2 Interface scattering rate calculation due to interface roughness on 2-D island size.

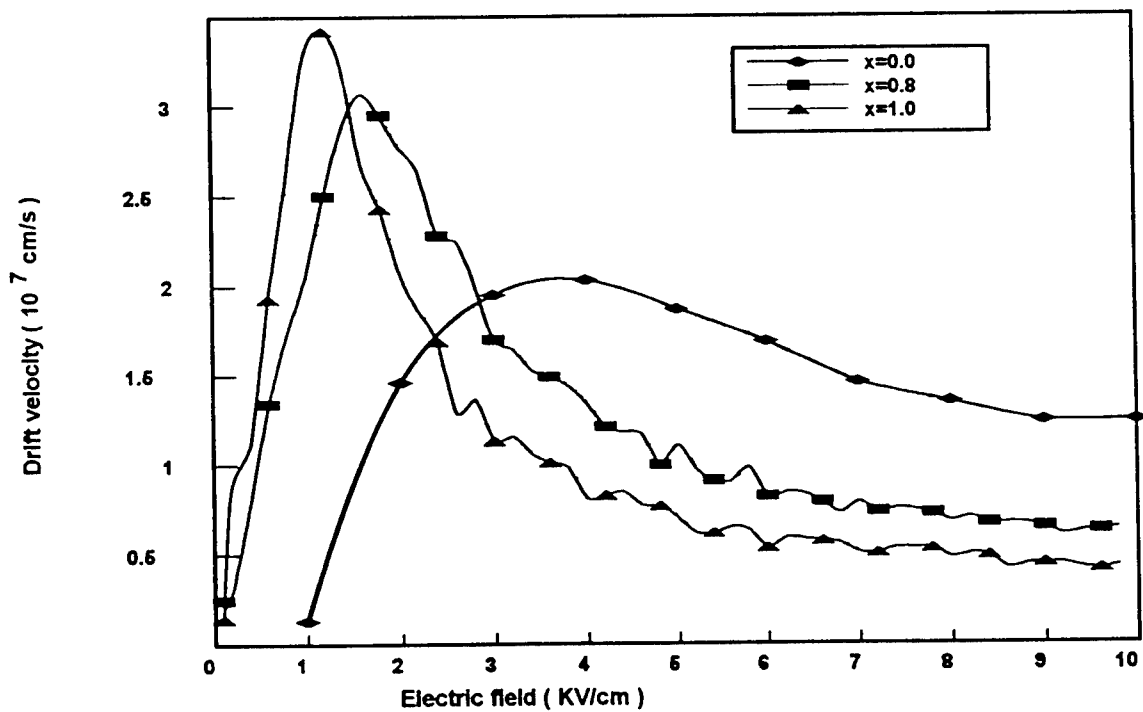


Fig.III.2-3 The drift velocity vs. electric field in $\text{In}_x\text{Ga}_{1-x}\text{As}$ as a function of In mole fraction

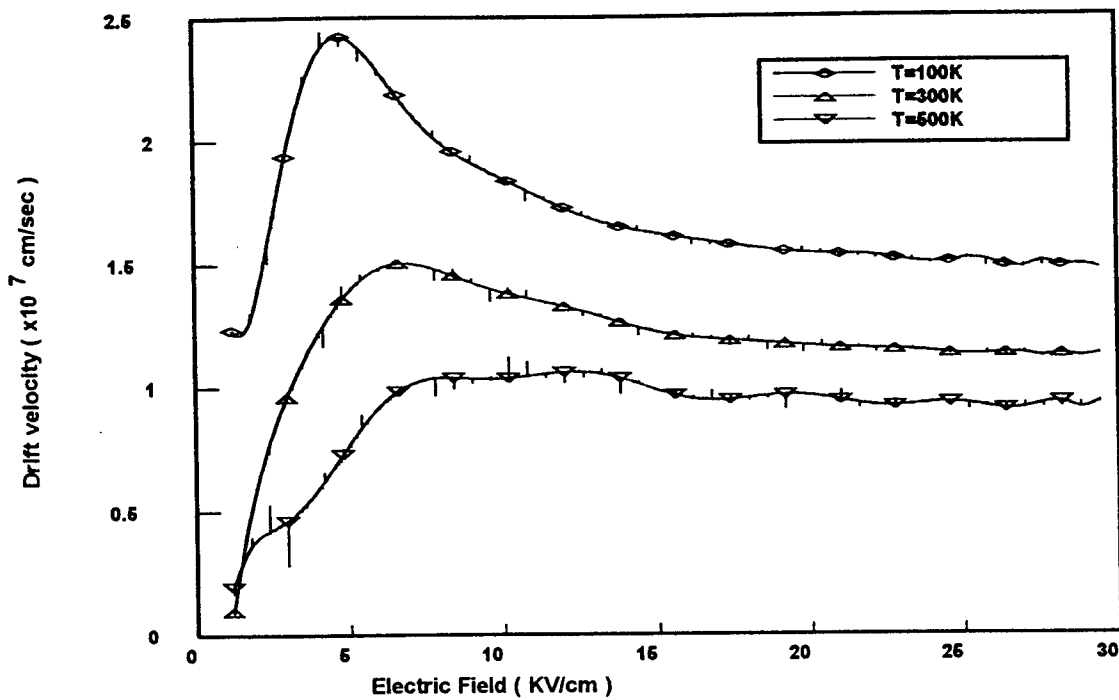


Fig.III.2-4 The drift velocity vs. electric field in $\text{Al}_x\text{Ga}_{1-x}\text{As}$ as a function of In mole fraction

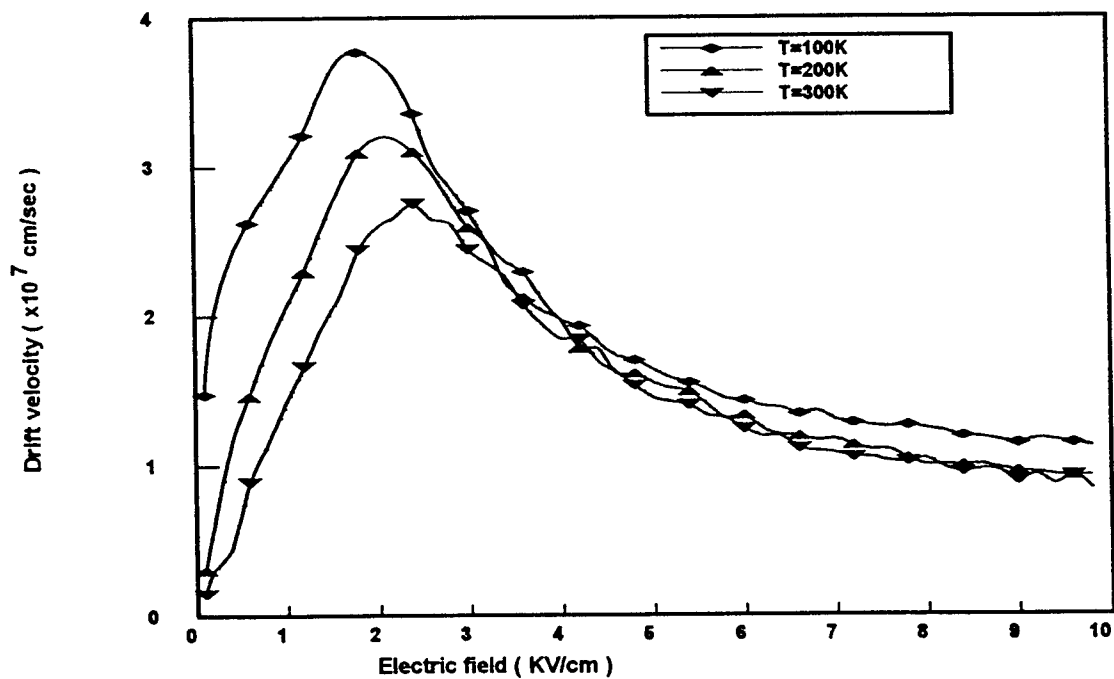


Fig.III.2-5 The drift velocity vs. electric field in $\text{In}_{0.53}\text{Ga}_{0.47}\text{As}$ at various temperatures

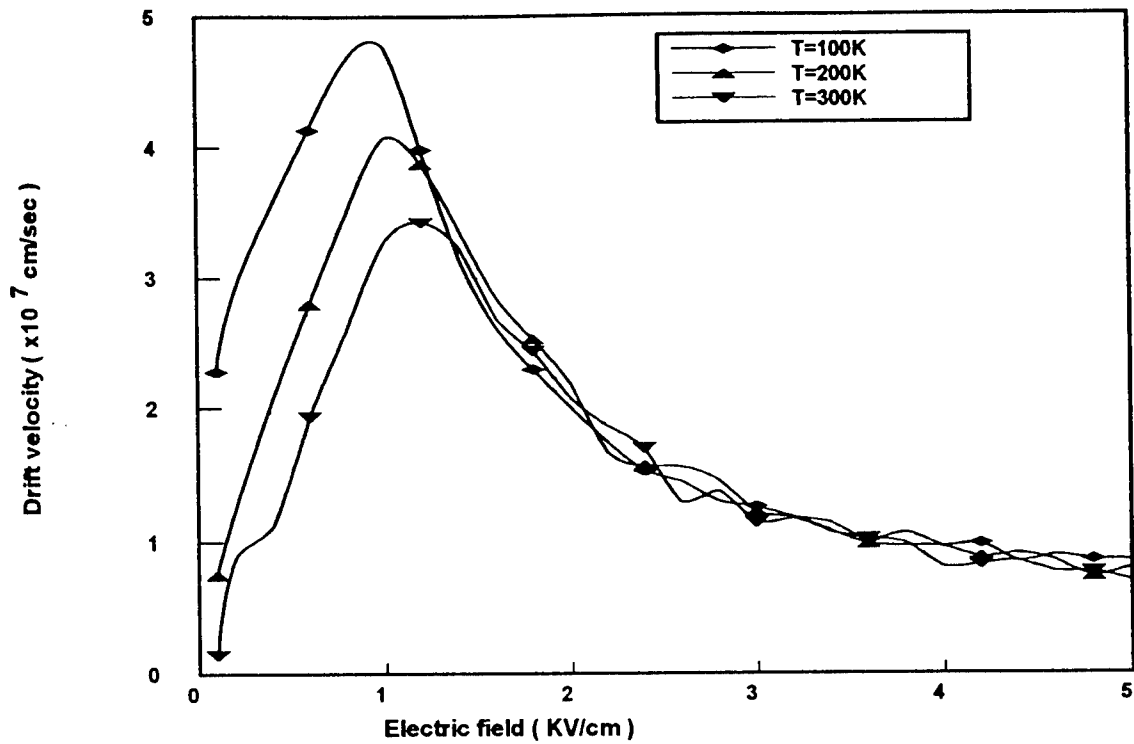


Fig.III.2-6 The drift velocity vs. electric field in InAs at various temperatures

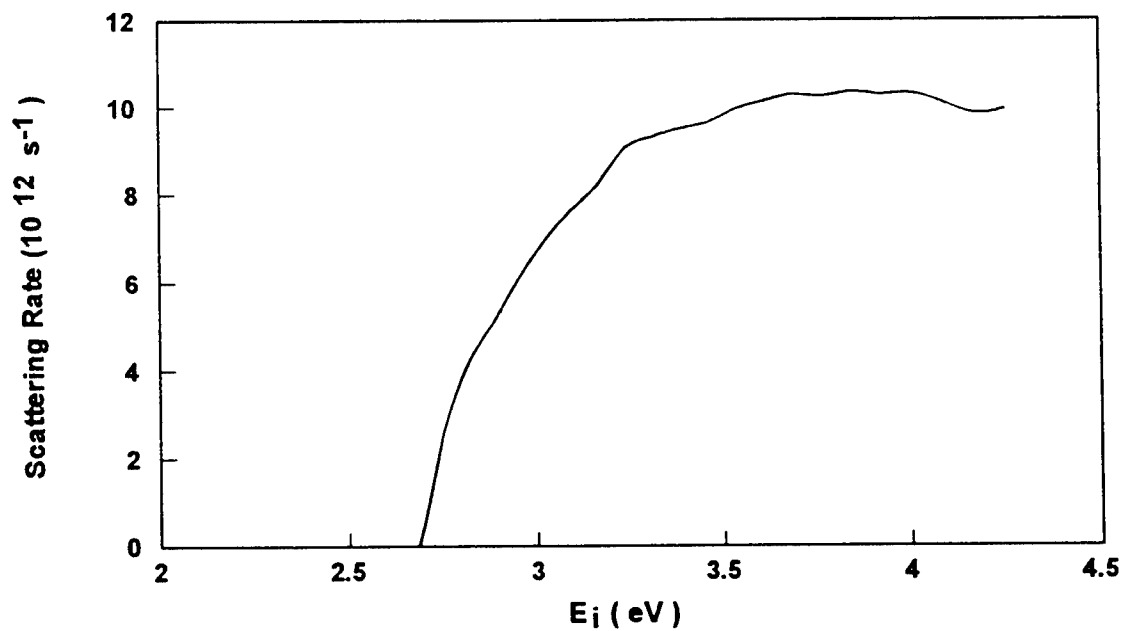


Fig. III.4-1. Impact ionization scattering rate calculation for InP

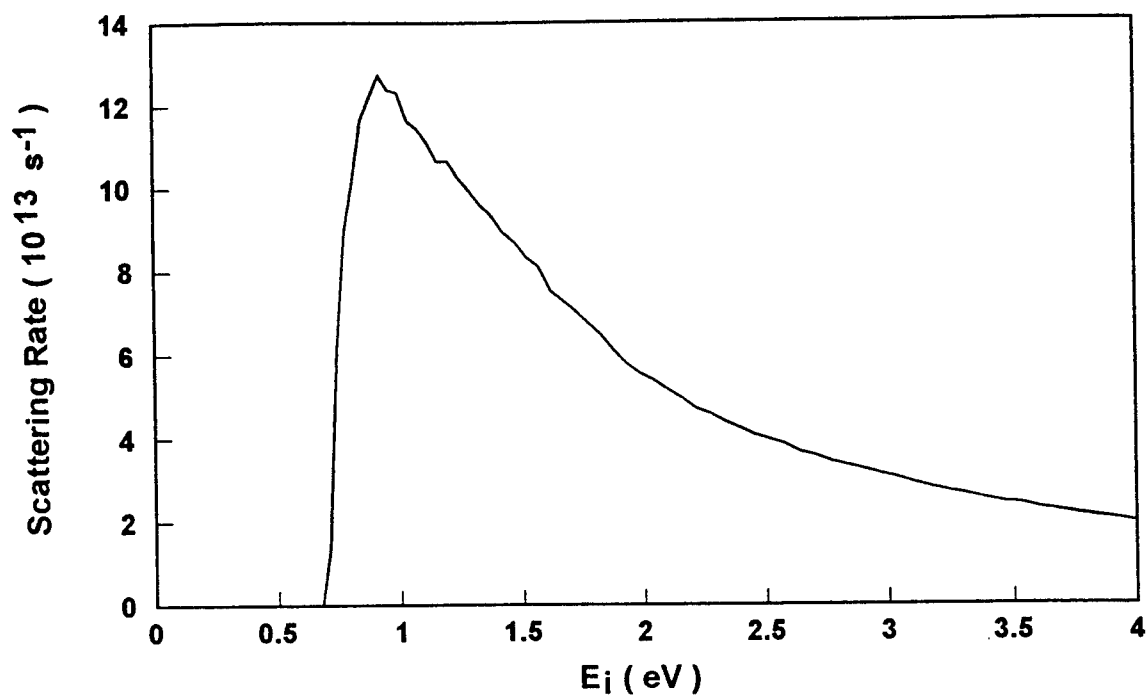


Fig. III.4-2. Impact ionization scattering rate calculation for InAs

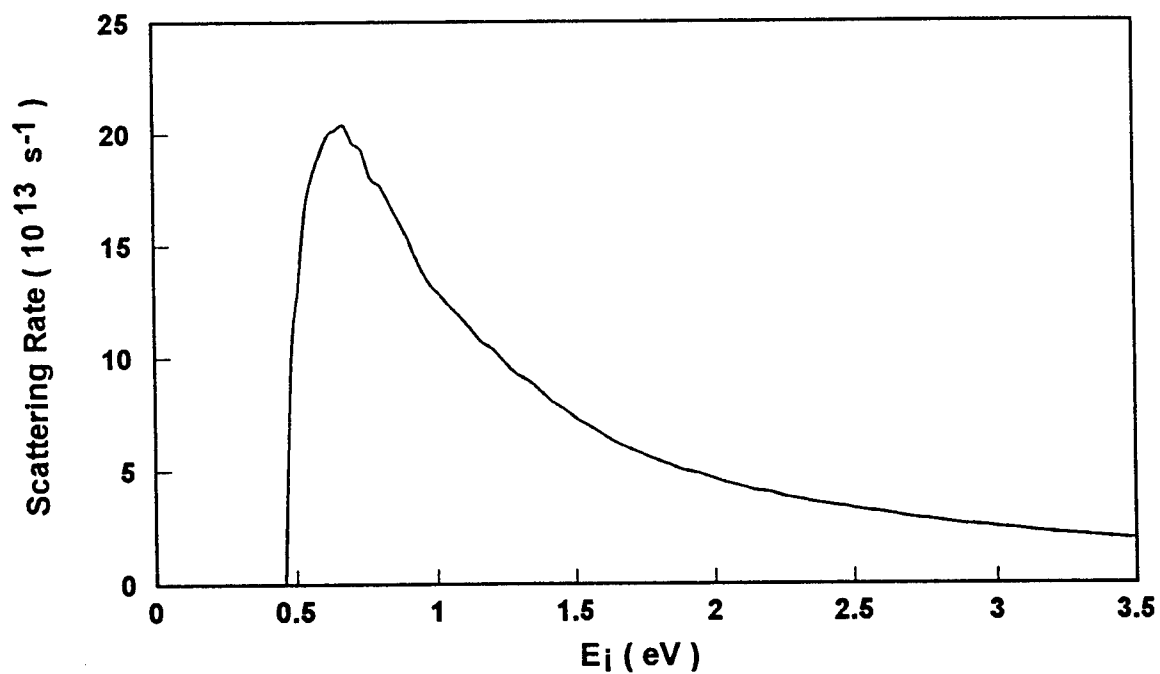


Fig. III.4-3. Impact ionization scattering rate calculation for InSb

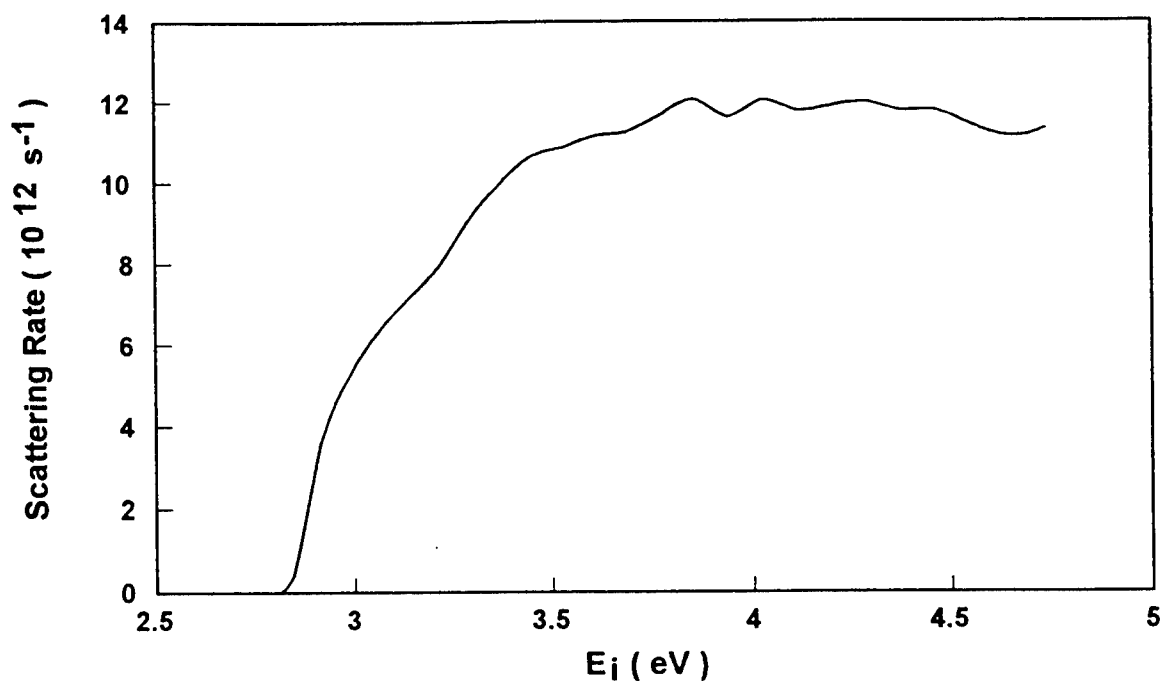


Fig. III.4-4. Impact ionization scattering rate calculation for GaAs

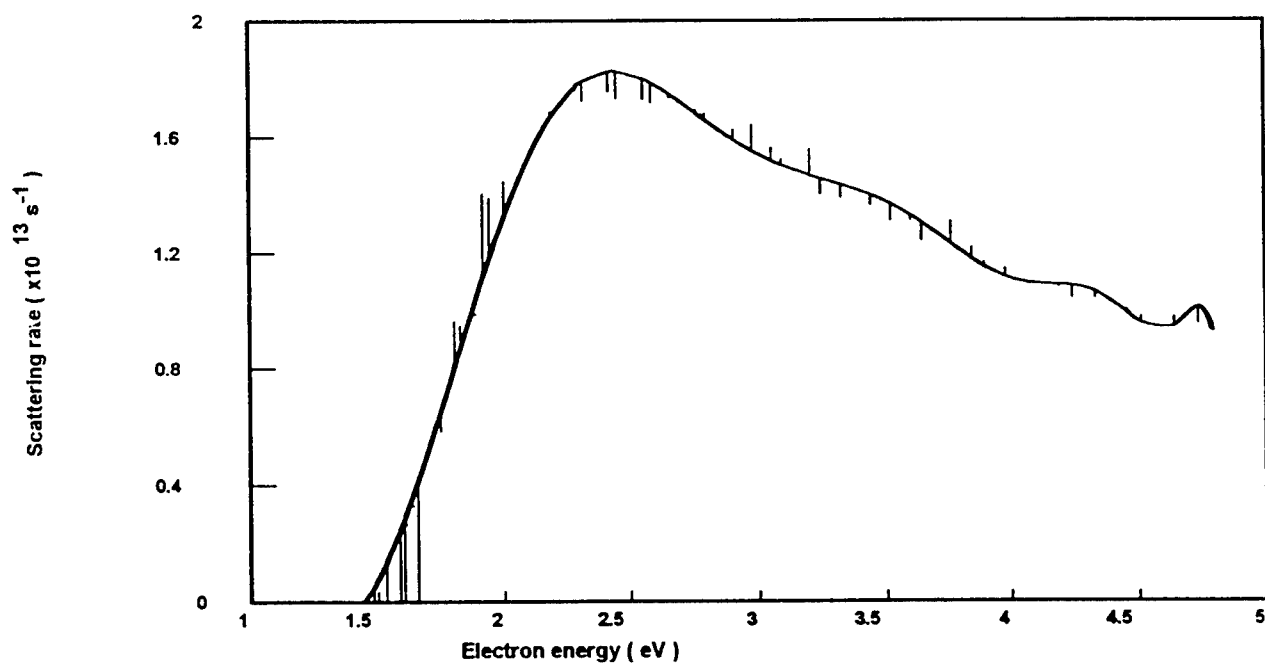


Fig. III.4-5 Impact ionization rate as a function of electron energy for $\text{In}_{0.53}\text{Ga}_{0.47}\text{As}$ at room temperature.

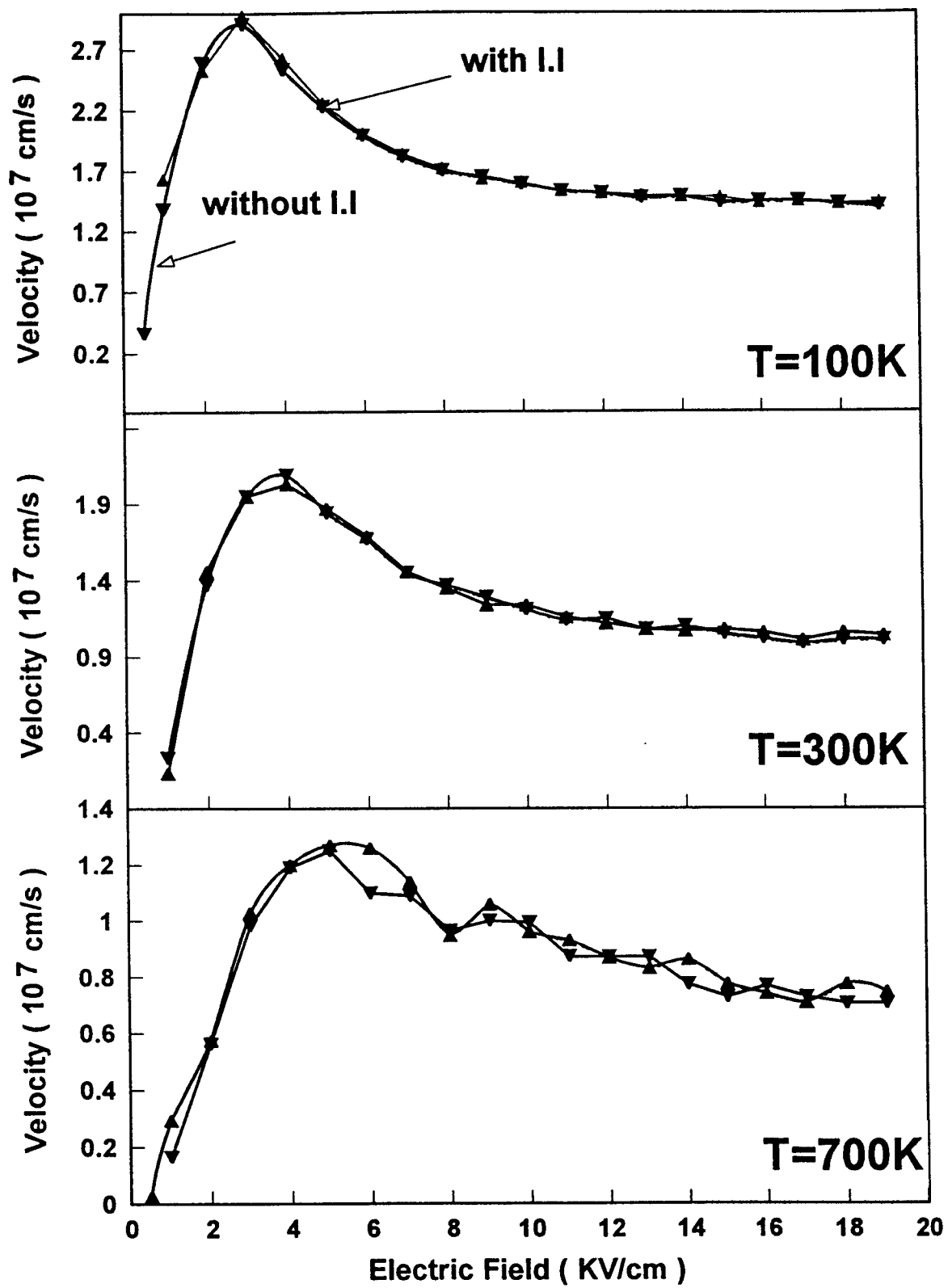


Fig.III.4-6 Drift velocity vs. electric field for GaAs at various temperatures

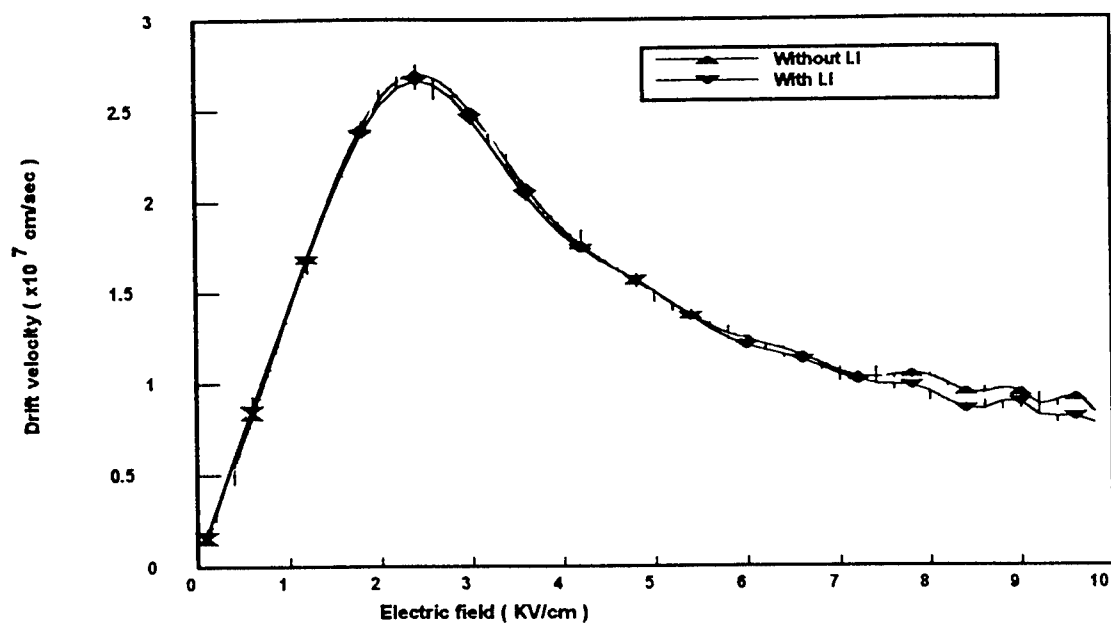


Fig.III.4-7 Drift velocity vs. electric field with and without impact ionization for $\text{In}_{0.53}\text{Ga}_{0.47}\text{As}$

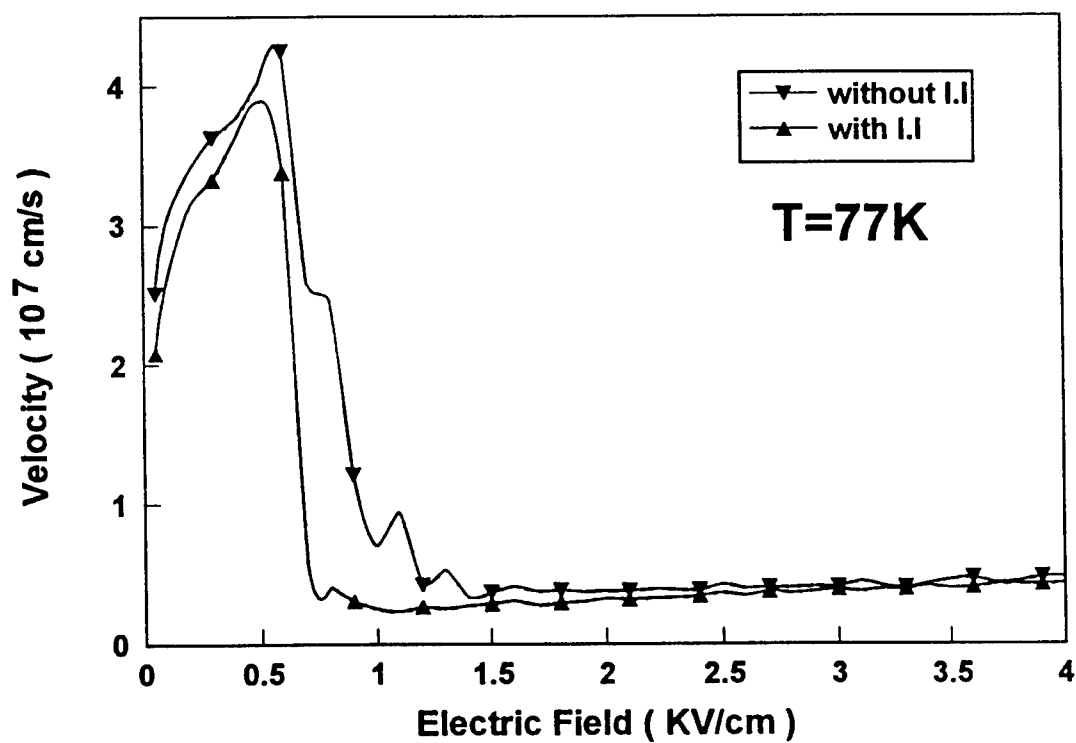


Fig.III.4-8 Drift velocity vs. electric field for InAs at 77K

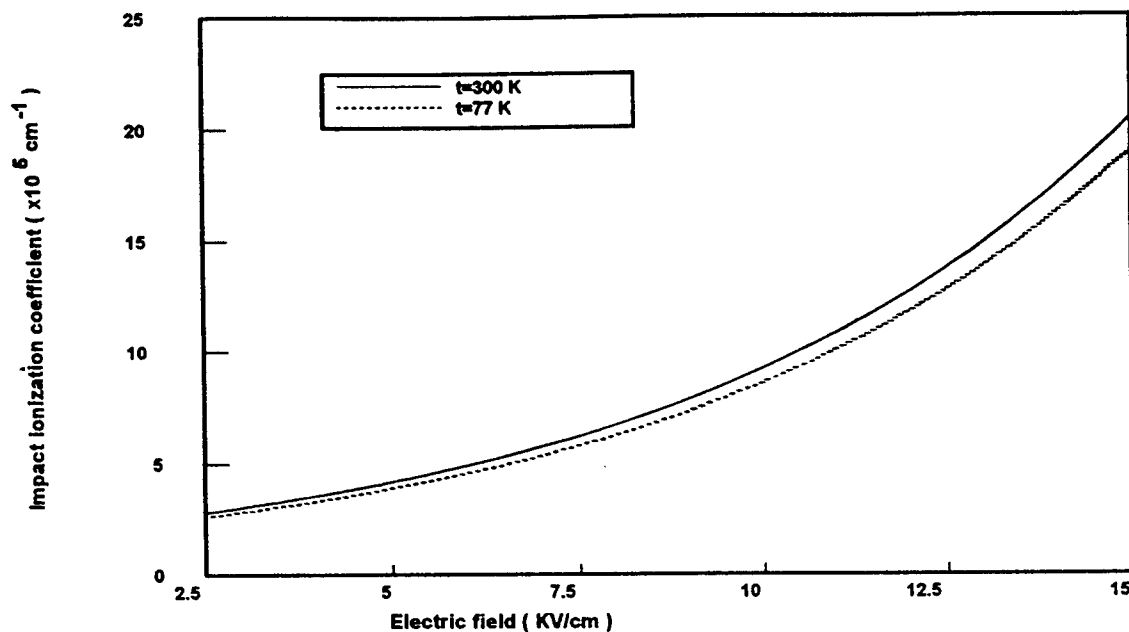


Fig. III.4-9 Impact ionization coefficient vs. electric field at different temperatures for $\text{In}_{0.53}\text{Ga}_{0.47}\text{As}$ at different temperature.

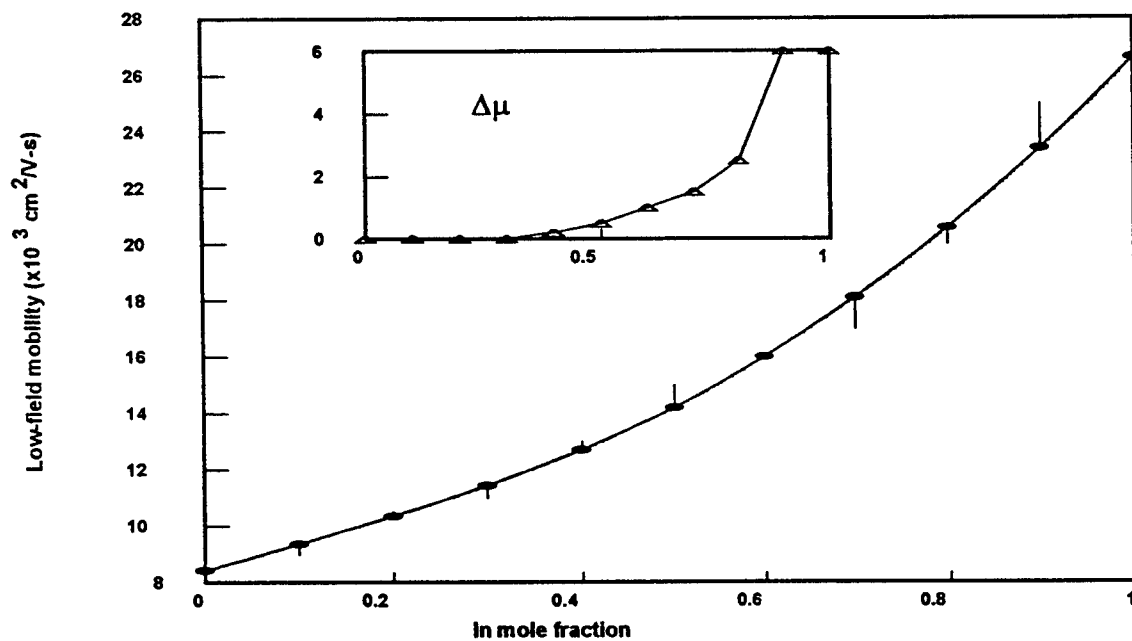


Fig. III.4-10 The low-field mobility in $\text{In}_x\text{Ga}_{1-x}\text{As}$ as a function of In mole fraction at room temperature (and the difference for μ between with and without impact ionization scattering).

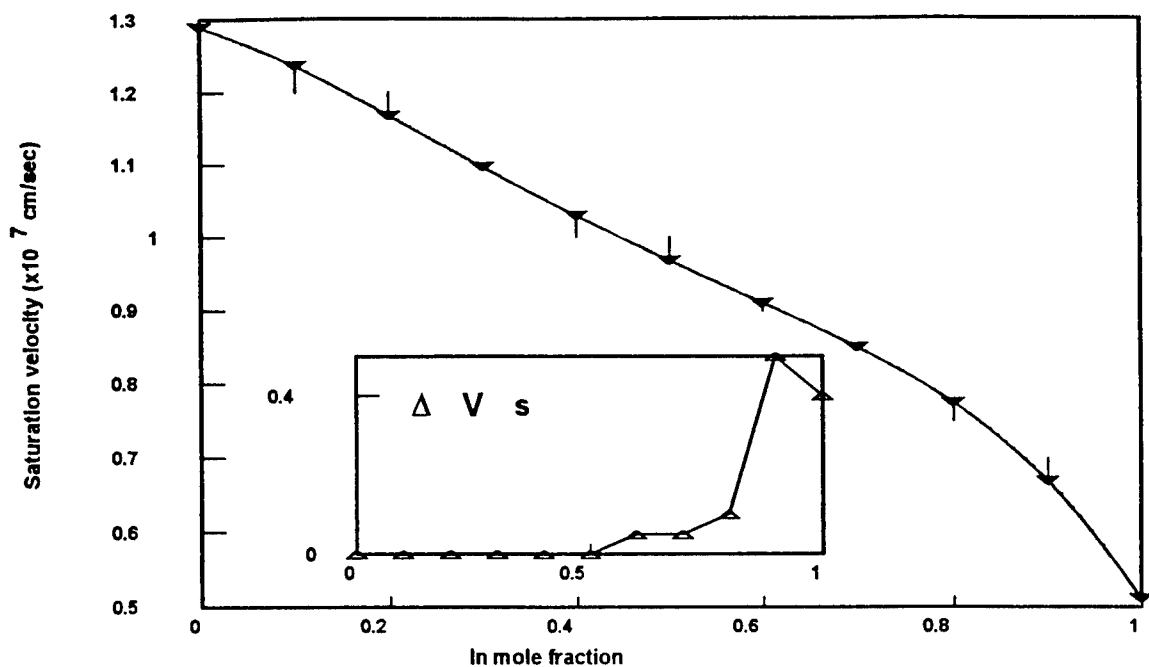


Fig. III.4-12 The saturation velocity in $\text{In}_x\text{Ga}_{1-x}\text{As}$ as a function of In mole fraction at room temperature (and the difference for V_s between with and without impact ionization scattering).

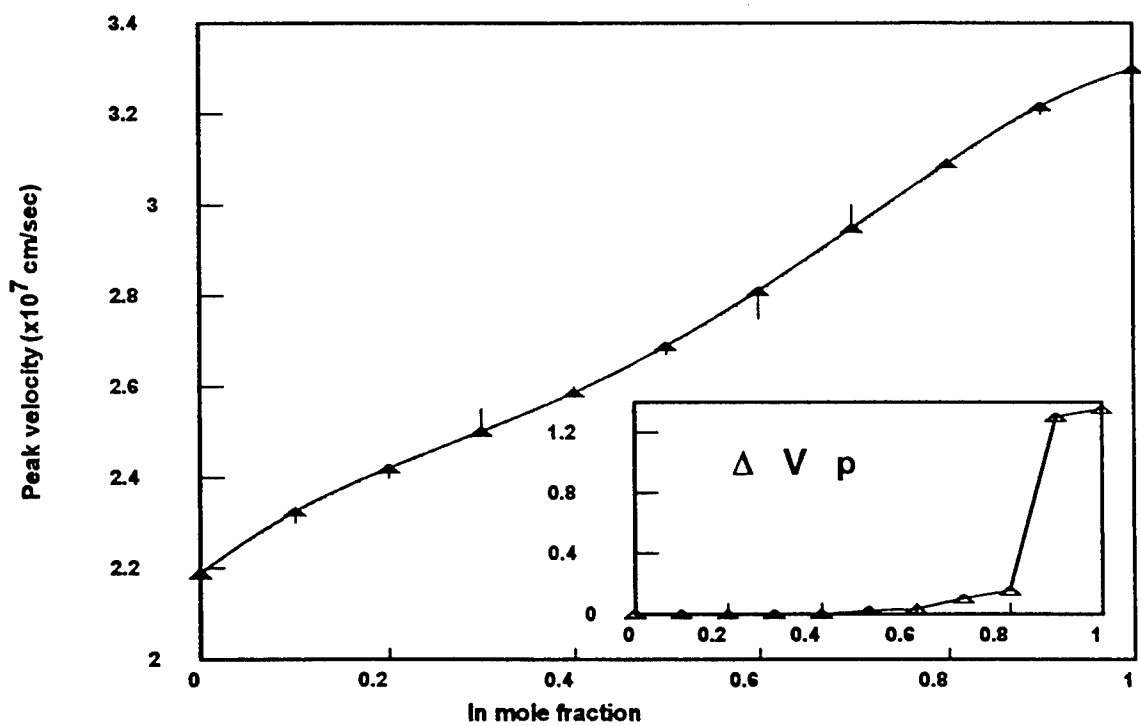


Fig. III.4-11 The peak velocity in $\text{In}_x\text{Ga}_{1-x}\text{As}$ as a function of In mole fraction at room temperature (and the difference for V_p between with and without impact ionization scattering).

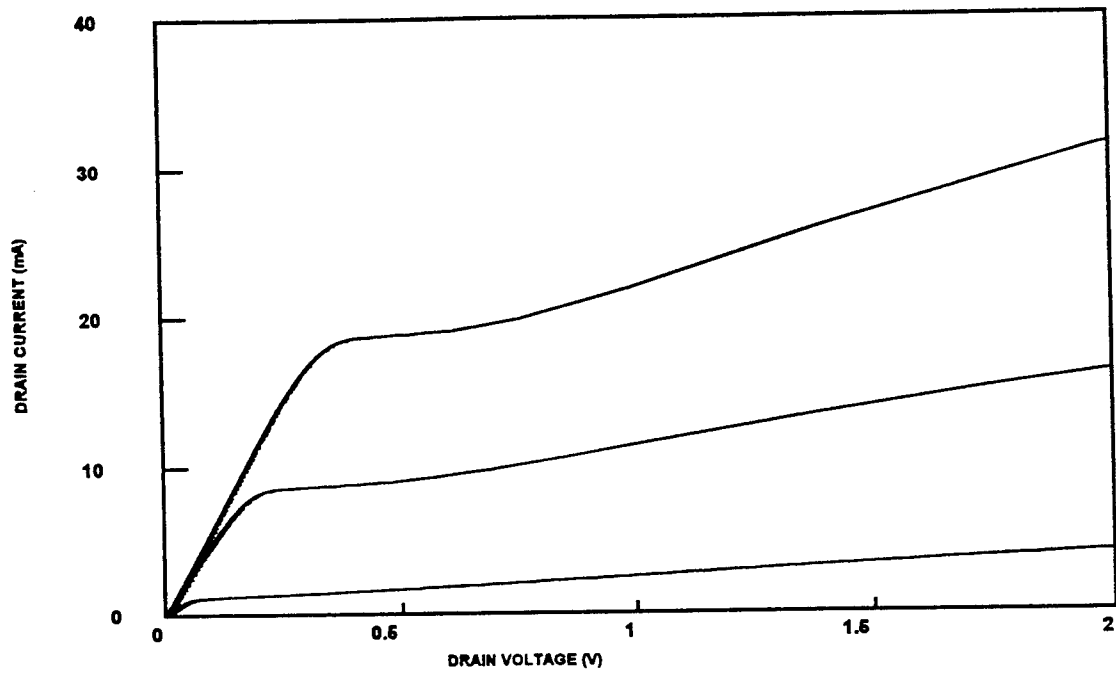


Fig. IV.1-1 Theoretical I-V curves.

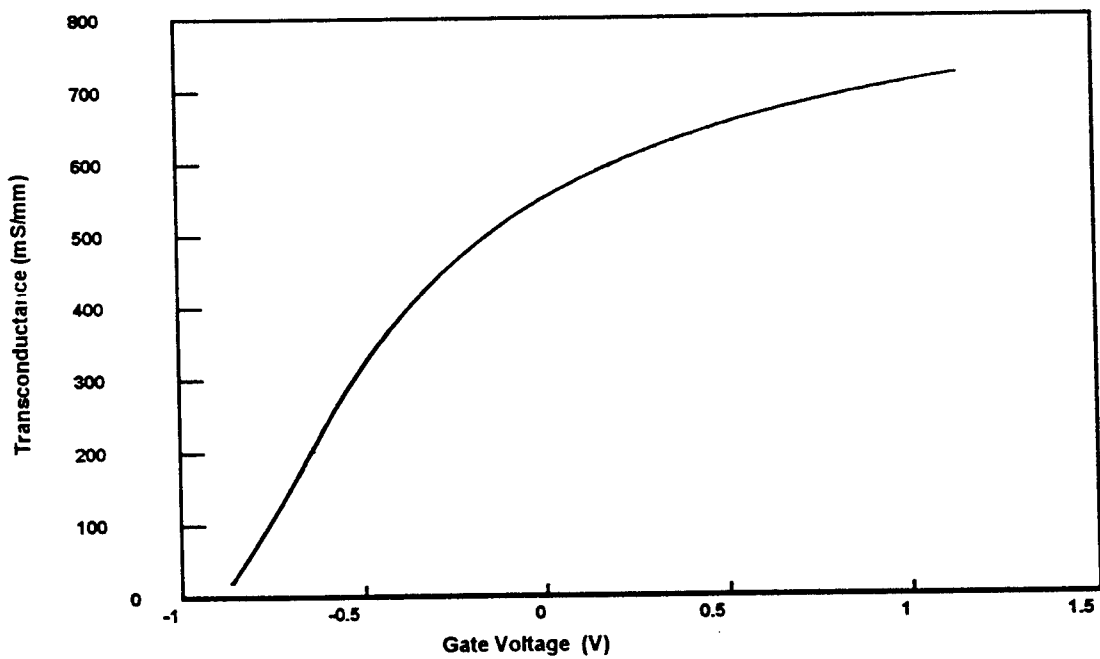


Fig. IV.1-2 Transconductance as a function of gate bias. $V_D=2.0V$.

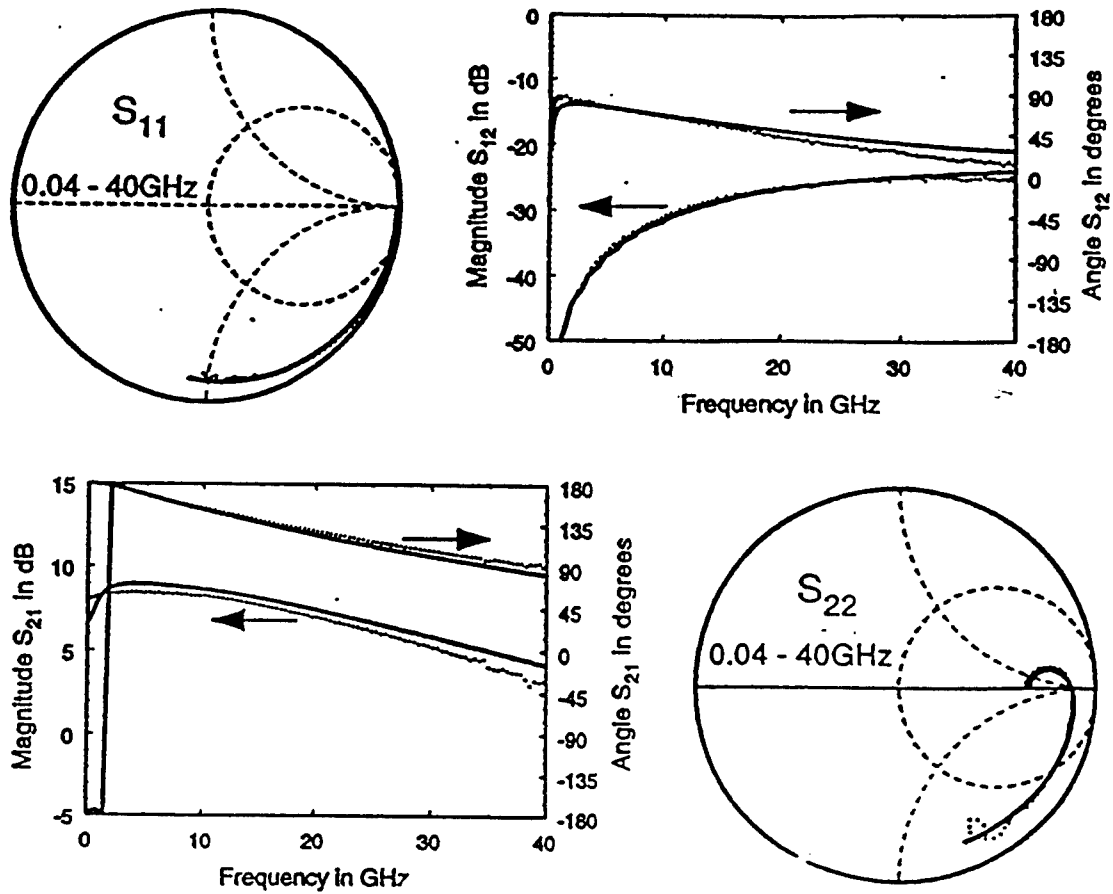


Fig.IV.2-1 Measured and modeled S-parameter of a 0.15 μ m \times 40 μ m In_{0.52}Al_{0.48}As/In_{0.53}Ga_{0.47}As HEMT.

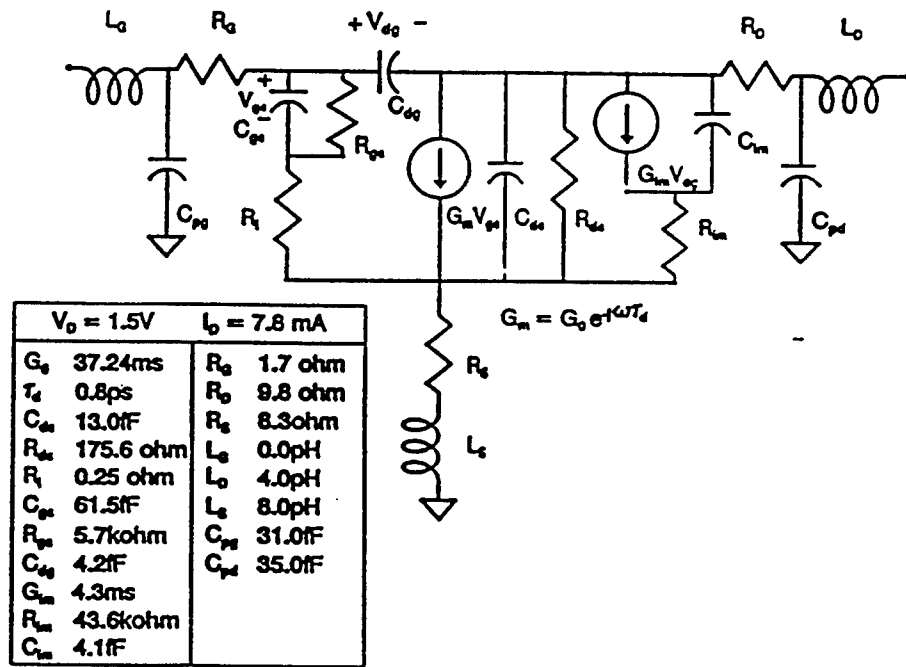


Fig.IV.2-2 Optimized small signal equivalent circuit.

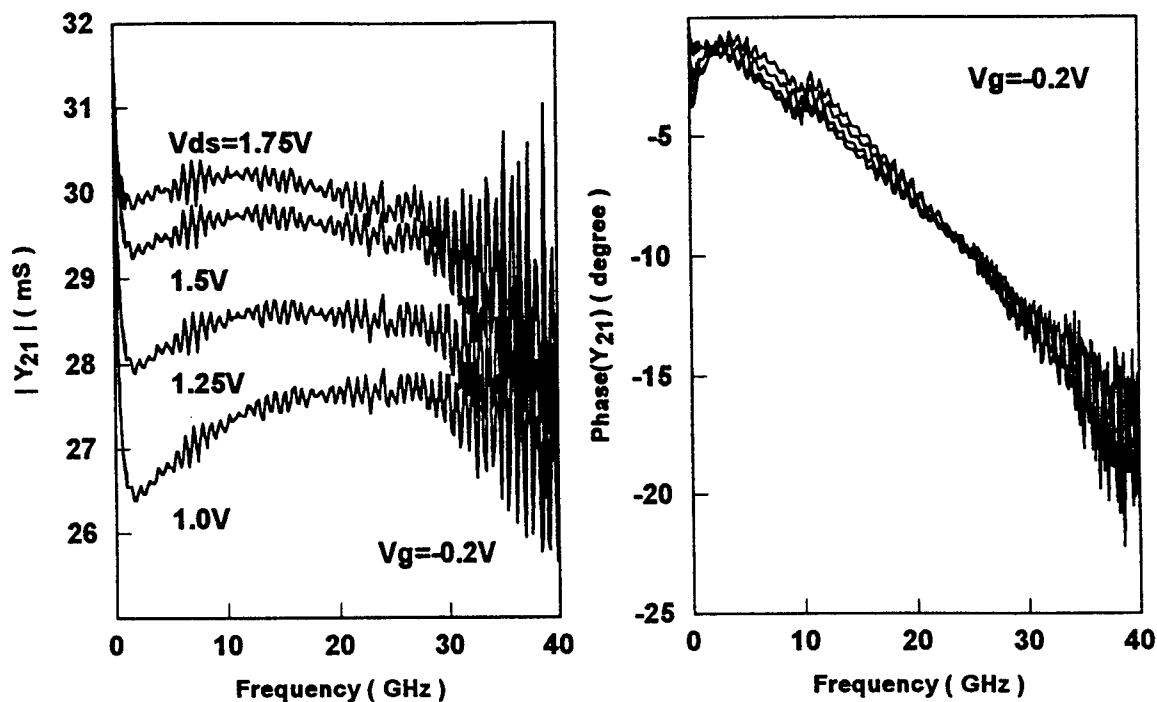


Fig.IV.2-3 Measured Y_{21} of a $0.15\mu\text{m} \times 40\mu\text{m}$ $\text{In}_{0.52}\text{Al}_{0.48}\text{As}/\text{In}_{0.53}\text{Ga}_{0.47}\text{As}$ HEMT as a function of source-drain bias for a gate bias of -0.2V.

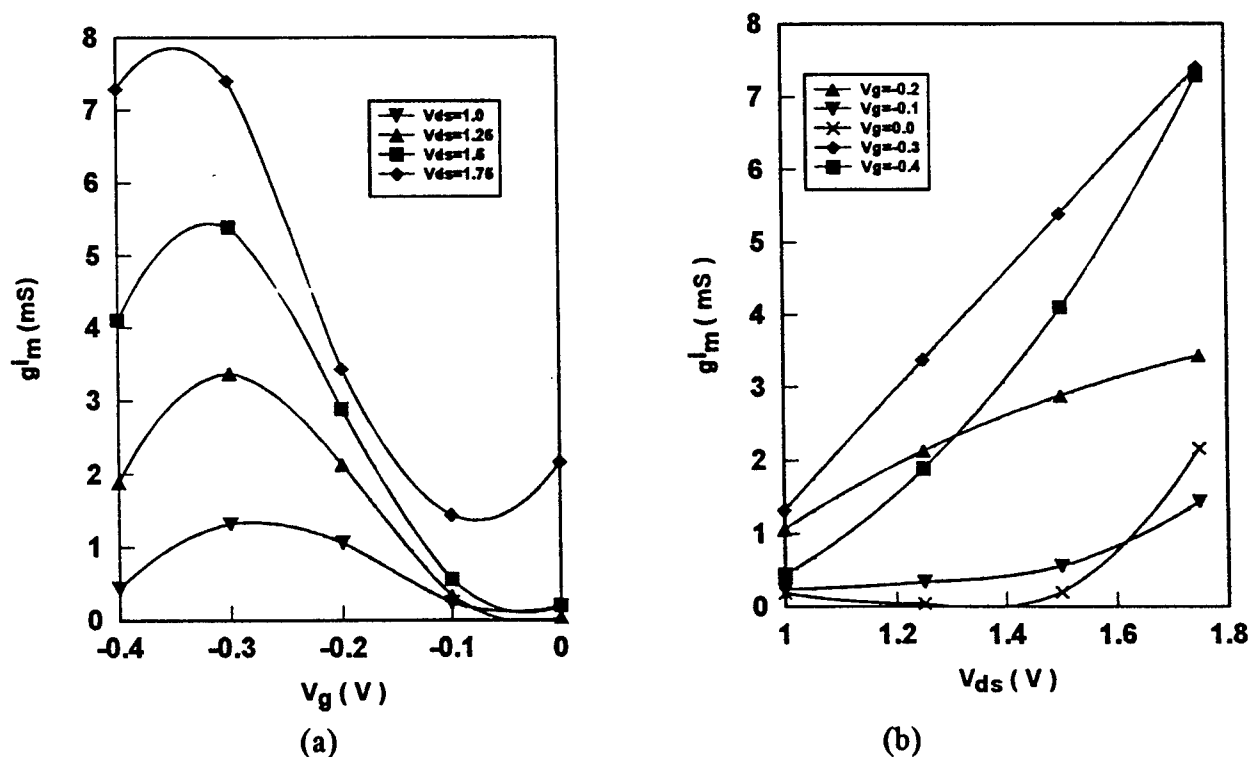


Fig. IV.2-4 Transconductance due to impact ionization in $\text{In}_{0.52}\text{Al}_{0.48}\text{As}/\text{In}_{0.53}\text{Ga}_{0.47}\text{As}$ HEMT

**INVESTIGATING THE ALGORITHMIC NATURE OF THE PROOF STRUCTURE
OF ORA LARCH/VHDL TO IMPROVE ITS PERFORMANCE**

**Ahmed E. Barbour
Professor of Computer Science
Mathematics and Computer Science Department**

**Georgia Southern University
P.O.Box 8093
Statesboro, GA 30460-8093**

**Final Report for:
Summer Research Extension Program
Rome Laboratory/ERDD**

**Sponsored by:
Air Force Office of Scientific Research
Bolling Air Force Base, DC**

and

Rome Laboratory/ERDD

December 1997

INVESTIGATING THE ALGORITHMIC NATURE OF THE PROOF STRUCTURE OF ORA LARCH/VHDL TO IMPROVE ITS PERFORMANCE

**Ahmed E. Barbour
Professor of Computer Science
Mathematics and Computer Science Department
Georgia Southern University**

Abstract

The algorithmic nature of the proof structure of Larch/VHDL theorem prover developed by Odyssey Research Associates (ORA) is investigated. General rules have been established to prove class of logic circuits with and without delay. The complexity of the prove for others classes are also investigated and suggestions to improve them are discussed. Several other issues related to improving the performance of the theorem prover and training scientists and engineers to understand the basic principles of the formal verification and the proof structure of the theorem prover as well as how to use Larch/VHDL theorem prover are also discussed.

INVESTIGATING THE ALGORITHMIC NATURE OF THE PROOF STRUCTURE OF ORA LARCH/VHDL TO IMPROVE ITS PERFORMANCE

Ahmed E. Barbour

1. Introduction

Hardware Description Language (VHDL) was established as an IEEE standard for the design and documentation of digital electronic systems. It was developed in response to the Computer Aided Design (CAD) community's need to handle larger and more complex designs and the need to be able to electronically exchange design information [1-4]. Simulation and formal verification are emerged as the only reasonable alternative tools which can be used effectively to verify complex systems. Simulation techniques use exhaustive testing of the VHDL model on several levels to determine the correctness of a design. Formal verification of hardware is a mathematical proof which shows that the design of a digital circuit satisfies certain properties regardless of the values of the inputs. Formal verification tools can also be used to prove that hardware designs satisfy properties such as functional correctness, security, and timing correctness.

Odyssey Research Associate (ORA) had developed an Ada verification environment, known as Penelope which also used in the development of Larch/VHDL theorem prover [5-8]. Penelope is based upon the Larch two-tiered specification language developed at the Massachusetts Institute of Technology (MIT) [10-13]. The first tier, the Larch Shared Language (LSL), is a first order predicate calculus used to build the traits, or theories, that define the sorts used by the target language (in the case of VHDL, types, such as bit, word, string, arrays, integer, .. etc.). The second tier, called the Interface Language, defines the communication mechanisms of the target language, in this case VHDL, in the Larch notation. LSL is used to mathematically model data objects and operations on those objects, while the interface language maps the VHDL model into the abstractions represented by the Larch expressions for the purpose of formal reasoning. Larch/VHDL verification environment is an interactive tool that helps its user to develop and verify digital electronic hardware designs written in VHDL [5-9]. The designer develops a VHDL model from a specification in a way that ensures the VHDL model will meet the specification. Figure 1 shows the structure of Larch/VHDL formal proof methodology.

This report investigates the algorithmic nature of the proof structure of Larch/VHDL theorem prover. Section 2 presents the general techniques used by Larch/VHDL to establish the proof of any logic structure expressed in VHDL. To understand the specification written for VHDL entities and the verification conditions that Larch/VHDL generates, Section 3 highlights some of the basic concepts of VHDL semantics in this area. Section 4 investigates the methodology used to establish the proof of different logic functions. General rules have been discovered and formulated for certain classes of logic functions in Section 5. Section 6 discusses some special cases in which the general rules are not applied. The importance of introducing formal methods into the undergraduate curriculum is presented in Section 7. Section 8 concludes the outcomes of this study and highlights the direction for further research works to be done in this field.

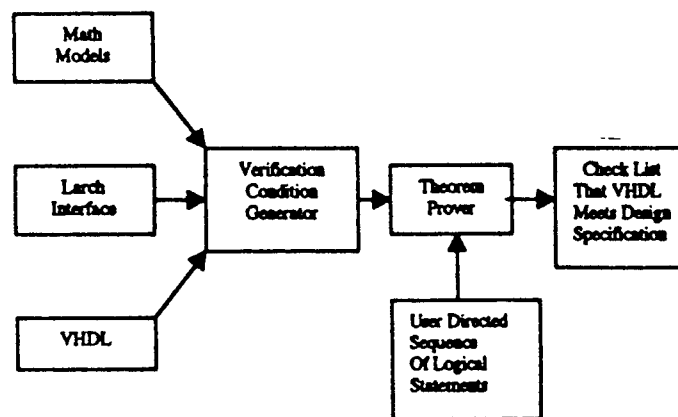


Figure 1: Formal Verification Using Larch/VHDL Technique

2. Techniques Used to Establish the Proof

Larch/VHDL verification system is a collection of tools integrated under the control of a graphical user interface called gui. The canvas of the gui depicts objects and relations. Larch/VHDL supports a window interface environment with several advanced features for entering specifications, developing code, and providing access to the Penelope theorem prover. The user creates an entity, writes its VHDL entity specification and writes its Larch/VHDL specification in the form of a list of guarantees. Then a command is invoked to install the Larch/VHDL specification in the library. An entity specification assigns to a VHDL entity three optional parts:

- (1) a set of imported theories,
- (2) a name and sort for its internal state, and
- (3) a list of guarantees, where a guarantee is a Boolean term in the language introduced by the imported theories [18].

Any syntax and sort-checking errors in the specification are reported by the system. Then the user creates a VHDL architecture for the entity and invokes the Verification Condition (VC) generator. It creates Larch theory that declare constants for all the ports, signals, and generics and asserts the constraints implied by the subtypes of these VHDL objects. The VC-theory contains an obligation to be proved of the form:

Given A, Prove That A Implies B ($A \rightarrow B$),

where A are the guarantees of the processes in the architecture body, and B are the guarantees from the specification of the entity being verified [8]. Once the system has generated the VC, the user creates a proof object and invokes the prover. While proving the VCs, the user can easily identify from the signal names which line of the VHDL architecture each hypothesis come from.

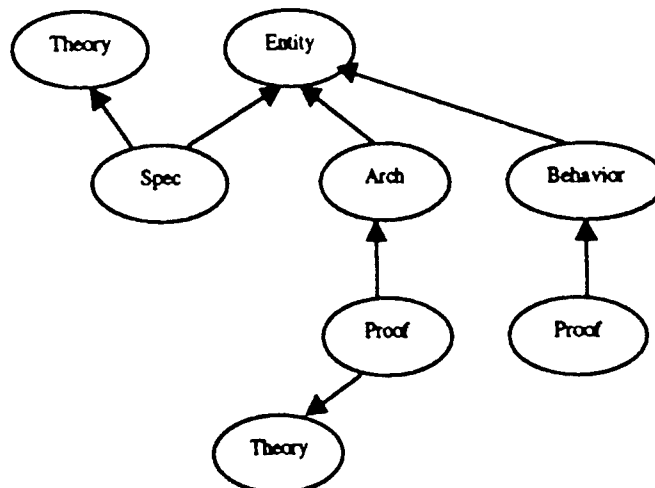


Figure 2: Larch/VHDL Proof Structure

If the user is unable to complete the proof because a particular hypothesis is not strong enough then the user knows which line of the VHDL is suspect. If all the proof obligations can be proved, then the given architecture satisfies the specification. All these proof steps will be illustrated by examples in Section 4. Figure 2 shows

the general relationship among entity, architecture, specifications, and theories used to help the proof construct.

The Larch/VHDL environment includes a large body of traits that define the basic constructs of digital design such as bit, vector, gate, logic operations and so on. Traits define sorts (logical types) and state properties or assertions that must hold true. Traits also contain theorems which are statements that are deducible from assertions, previously deduced theorems, and/or the assertions or theorems of other traits that are included. The two-tiered Larch approach allows designers the capability to extend the library of traits in order to support user defined sorts in their models. Once implemented, the traits are available as library components for reuse in other applications. Traits are used to capture the concepts and relationships used in digital design. There are traits devoted to arithmetic concepts, and to data structures such as arrays and lists. To support VHDL semantics there are traits defining signals, and signal delay, and other concepts needed to express the semantics of VHDL. There are traits that describe the relationship between bit level operations and their arithmetic interpretation, in 2's complement or unsigned bit-level representations.

Penelope Larch/VHDL theorem prover includes a simple proof editor/checker for predicate calculus that provides a number of proof rules for performing simplification and proofs. Penelope applies the rules according to user directions and indicates to the user what, if anything, still has to be proved after each step. Each statement to be proved or simplified is presented in the form of a sequent, a set of hypotheses and a conclusion. Each proof in Penelope takes place in the context of an available theory. Within a VHDL design unit, the theory is determined by entity declaration annotations and all the local lemmas currently being applied to complete the proof. The theory that is available for proving a given lemma consists of the axioms, assumptions, and proved lemmas that precede the given lemma.

3. Some of The Basic Concepts of VHDL Semantics

The basic concepts of the semantics of the VHDL as described in the Language Reference Manual (LRM) and how they are modeled in Larch/VHDL are very important to understand the specifications of the VHDL entities and the VCs generated by the system. The semantics of VHDL describe what a VHDL simulator must do. Briefly, a simulator will start by elaborating the VHDL code to create processes, signals and their drivers and initializing them. Then, the simulator proceeds by executing simulation cycles consecutively, updating the processes signals and drivers and advancing time at each cycle [18]. In order to write specification, terminology

- (4) proof-structuring rules, such as proof by cases or proof by induction.

The following examples illustrate the components of the proof structure of a VHDL unit and the commands used to perform the required proof.

5.1 Simple Logic Gate Example: Two-Input AND Gate

Figure 4 shows the two-input AND gate and its proof structure as created by Larch/VHDL.

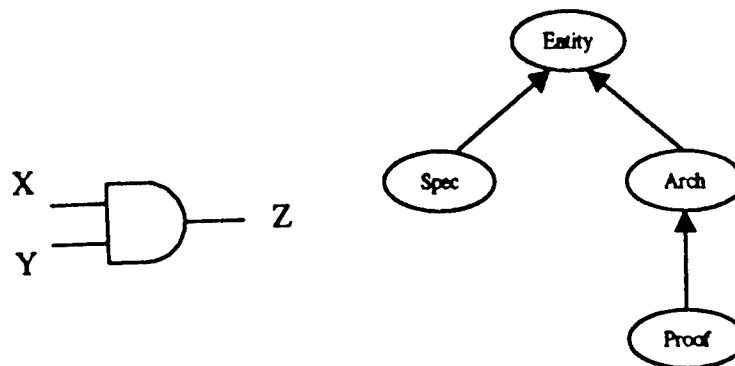


Figure 4: Two-Input AND Gate and Its Proof Structure

This example illustrates the proof methodology for a simple logic gate. All other logic gates follow the same proof technique established in this example.

- (1) The proof structure for any logic gate is generated by creating a VHDL entity to the AND gate as shown.

```

entity and2 is
  port ( x, y : Bit; z : out Bit);
end and2;
  
```

Invoking the command: **Make Lib** from the entity icon will check the correctness of the VHDL syntax and if it is correct, will create a library for the AND gate.

- (2) The second step is to create the arch icon from the entity icon as shown.

```

architecture arch of and2 is
begin
  z <= x and y;
end arch;
  
```

- (3) The third step is to express specifications in a specification language provided by Larch/VHDL system. It provides a specification language for asserting specifications of functionality and timing. The created specifications from the entity icon is shown.

```
entity and2
  guarantees
  always
  ( z = '1' ) = ( ( x and y ) = '1' )
  with z delayed by 0 from x, y
end
```

Invoking the command: Make Lib from the specification icon will check the syntax correctness of the specification and if no error is generated, it will create a specification Library.

- (4) The last step is to create the proof icon and invokes the Verification Condition (VC) generator from the architecture icon by using the command: Generate Verification which generates the VCs required to be proved by the Penelope. If there is no error generated by all these three steps, the prover may be invoked from the proof icon. The proof screen will open asking to add the VCs to be proved. If you click on the command: Insert Obligation, the VCs will be inserted in its position on the screen as shown below.

```
--> trait arch_and2_VC has unfinished proofs
--l Larch
proof section for arch_and2_VC
--l proof section
--> trait arch_and2_VC has unfinished proofs
--l vc in trait arch_and2_VC
--l proof:
  >> (Forall a:
    State::
      y'last_event @ a > 0 and x'last_event @ a > 0
      -> z @ a = (x @ a and y @ a)
      and z'last_event @ a >= min(y'last_event @ a, x'last_event @ a))
  -> (Forall a:
    State::
      x'last_event @ a > 0 and y'last_event @ a > 0
      -> (z @ a = '1') = ((x @ a and y @ a) = '1')
      and z'last_event @ a >= min(x'last_event @ a, y'last_event @ a))
  <proof>
--l end proof section
--l end Larch
```

It is clear that the VHDL code of two-input AND gate shown in the architecture part is converted to Larch statements in the form:

>> Forall a (time), if y is stable and x is stable for $a > 0$,
 -> it implies that $z \text{ at } a = x \text{ at } a \text{ and } y \text{ at } a$
 and z will have been stable at least since the most recent change in x and y.

Also, the specification code of the AND gate is converted to Larch statements in the form similar to the architecture form shown previously. The proof may be constructed using the simple logical argument:

Given A, Prove that A Implies B.

The Penelope environment provides an opportunity to record and check the designer's reasoning. In almost all cases, the proof problem presented is partitioned and transformed using previously proven theorems. The step-by-step proof of two-input AND gate is as shown.

- (1) The first step is to strip off the quantifier and make the precondition into hypothesis and do simplification by using the command: BY synthesis/analysis of FORALL/IMPLIES as shown.

```

--| proof section
--> trait arch_and2_VC has unfinished proofs
--| vc in trait arch_and2_VC
--| proof:
  BY synthesis/analysis of FORALL/IMPLIES
  1. y'last_event @ a > 0 and x'last_event @ a > 0
  -> z @ a = (x @ a and y @ a)
    and z'last_event @ a >= min(y'last_event @ a, x'last_event @ a)
  2. Forall a:
  State::
  y'last_event @ a > 0 and x'last_event @ a > 0
  -> z @ a = (x @ a and y @ a)
    and z'last_event @ a >= min(y'last_event @ a, x'last_event @ a)
  3. x'last_event @ a > 0
  4. y'last_event @ a > 0
  >> (z @ a = '1') = ((x @ a and y @ a) = '1')
  and z'last_event @ a >= min(x'last_event @ a, y'last_event @ a)
  <proof>
--| end proof section
--| end Larch

```

The command: synthesis-conclusion, is first selected and then the second selection is the command:

Forall/implies synthesis, which is very often used as the first step in a proof. If the goal is of the form

Forall alpha A -> B,

the forall quantifier is peeled off and the implies antecedent, A, is made a hypothesis. (alpha is the time variable which considers as a free variable (a).) If we add: with analysis, as an option for the previously selected forall synthesis, it invokes forall analysis and applies it on the result. As shown from the example, the for all quantifier has reduced to simple four logical imply statements which need to be proven.

- (2) The second step is to apply the command: `simplify` and select the option: SDVS simplification as shown.

```
--l proof section
--> trait arch_and2_VC has unfinished proofs
--l vc in trait arch_and2_VC
--l proof:
  BY synthesis/analysis of FORALL/IMPLIES
  BY simplification
  1. 0 < y'last_event @ a
  2. 0 < x'last_event @ a
  3. z @ a = (x @ a and y @ a)
  4. min(y'last_event @ a, x'last_event @ a) <= z'last_event @ a
  5. Forall a:
    State::
      0 < y'last_event @ a and 0 < x'last_event @ a
      -> z @ a = (x @ a and y @ a)
          and min(y'last_event @ a, x'last_event @ a) <= z'last_event @ a
      >> min(x'last_event @ a, y'last_event @ a) <= z'last_event @ a
      <proof>
--l end proof section
--l end Larch
```

The command: `simplify` has several kinds of simplification. In some cases a series of simplifications is useful. One powerful simplification tool is SDVS simplify. It is a decision procedure which can reduce valid formulas about the integers with "+", "<", "<=", to true. That means, it is effectively applied to simple arithmetic, inequalities, predicate-free expressions. It is very useful near the end of a proof. It is clear that this type of simplification is not powerful enough to complete the proof.

- (3) The final step is to use the command: `add +` which applies rewriting rule to our simplification. So when the step `By simplification+` occurs, the goal is simplified by using all the equations marked `morphism` as left to right rewrite rules. For example, if `n, m, k` are variables of sort `Nat`, the goal

```
(n <= m and m <= k) --> (n <= k) would rewrite to:
(int(n) <= int(m) and int(m) <= int(k)) --> (int(n) <= int(k))
```

and this would simplify to true. This procedure is shown below.

```
proof section for arch_and2_VC
--l proof section
--l vc in trait arch_and2_VC
--l proof:
  BY synthesis/analysis of FORALL/IMPLIES
  BY simplification+
  BY synthesis of TRUE
--l end proof section
```

--| end Larch

This last step proves the VCs by issuing the final command: **By synthesis of TRUE**.

It is clear that this proof methodology is short and very powerful. The proof could be lengthy if we select another methodology. For example, using the following commands:

- (1) **synthesis-conclusion** followed by **Forall/implies synthesis** will produce similar result as shown in step one.
- (2) If we use the command: **synthesis-conclusion** again and then selecting the command: **and synthesis**, the result of the proof simplification will be as shown.

```
--| proof:
  BY synthesis of FORALL/IMPLIES
  BY synthesis of AND
  1. a:
  State::
  y'last_event @ a > 0 and x'last_event @ a > 0
  -> z @ a = (x @ a and y @ a)
    and z'last_event @ a >= min(y'last_event @ a, x'last_event @ a)
  2. x'last_event @ a > 0
  3. y'last_event @ a > 0
  >> (z @ a = '1') = ((x @ a and y @ a) = '1')
  <proof>
  1. Forall a:
  State::
  y'last_event @ a > 0 and x'last_event @ a > 0
  -> z @ a = (x @ a and y @ a)
    and z'last_event @ a >= min(y'last_event @ a, x'last_event @ a)
  2. x'last_event @ a > 0
  3. y'last_event @ a > 0
  >> z'last_event @ a >= min(x'last_event @ a, y'last_event @ a)
  <proof>
--| end proof section
--| end Larch
```

The command: **and-synthesis** is used when the goal has the form

>> A and B.

This step splits the proof into two subproof, one with goal A, and the other with goal B. If the conjunction is larger, then more subproofs are created. Now we have two subproofs which need to be proved.

- (3) The next step is to prove the first subproof as shown.

```
--| proof:
  BY synthesis of FORALL/IMPLIES
```

BY synthesis of AND
 BY analysis of FORALL,
 WITH a FOR a:State

1. Forall a:

State::

```
y'last_event @ a > 0 and x'last_event @ a > 0
-> z @ a = (x @ a and y @ a)
   and z'last_event @ a >= min(y'last_event @ a, x'last_event @ a)
2. x'last_event @ a > 0
3. y'last_event @ a > 0
4. y'last_event @ a > 0 and x'last_event @ a > 0
-> z @ a = (x @ a and y @ a)
   and z'last_event @ a >= min(y'last_event @ a, x'last_event @ a)
  >> (z @ a = '1') = ((x @ a and y @ a) = '1')
<proof>
```

1. Forall a:

State::

```
y'last_event @ a > 0 and x'last_event @ a > 0
-> z @ a = (x @ a and y @ a)
   and z'last_event @ a >= min(y'last_event @ a, x'last_event @ a)
2. x'last_event @ a > 0
3. y'last_event @ a > 0
  >> z'last_event @ a >= min(x'last_event @ a, y'last_event @ a)
<proof>
```

--| end proof section

--| end Larch

Using the command: analyze-hypothesis, it will open to you the sub-menu buttons which contain several forms of the selected hypothesis. These hypothesis options are:

and-analysis, equals-analysis, exists-analysis, forall-analysis,

if-else-analysis, if-then-analysis, imp-analysis, not-equals-analysis,

not-analysis, or-analysis, xor-analysis, and-analysis, ...etc.

Our selection will be imp-analysis which will simplify the first subproof as shown above.

- (4) The next step is to use the command: thinning which is used to thin out hypotheses that are not needed to complete the proof. There are a variety of template matching parameters to help describe the thinning rule to be applied. We will use one of them, the command: (binding *) as shown.

--| proof:

BY synthesis of FORALL/IMPLIES

BY synthesis of AND

BY analysis of FORALL,

WITH ga FOR ga:State

BY thinning (binding *)

1. x'last_event @ a > 0

2. y'last_event @ a > 0

```

3. y'last_event @ a > 0 and x'last_event @ a > 0
-> z @ a = (x @ a and y @ a)
and z'last_event @ a >= min(y'last_event @ a, x'last_event @ a)
>> (z @ a = '1') = ((x @ a and y @ a) = '1')
<proof>
1. Forall a:
State::
y'last_event @ a > 0 and x'last_event @ a > 0
-> z @ a = (x @ a and y @ a)
and z'last_event @ a >= min(y'last_event @ a, x'last_event @ a)
2. x'last_event @ a > 0
3. y'last_event @ a > 0
>> z'last_event @ a >= min(x'last_event @ a, y'last_event @ a)
<proof>
--l end proof section
--l end Larch

```

- (5) The last step of this subproof is to use the command: **simplify with SDVS simplification option** to complete the proof as shown.

```

--l proof:
BY synthesis of FORALL/IMPLIES
BY synthesis of AND
BY analysis of FORALL,
WITH a FOR a:State
BY thinning (binding *)
BY simplification
BY synthesis of TRUE
1. Forall a:
State::
y'last_event @ a > 0 and x'last_event @ a > 0
-> z @ a = (x @ a and y @ a)
and z'last_event @ a >= min(y'last_event @ a, x'last_event @ a)
2. x'last_event @ a > 0
3. y'last_event @ a > 0
>> z'last_event @ a >= min(x'last_event @ a, y'last_event @ a)
<proof>
--l end proof section
--l end Larch

```

- (6) We repeat the same proof steps of the first subproof to prove the next subproof as shown.

```

--l proof:
BY synthesis of FORALL/IMPLIES
BY synthesis of AND
--l Subproof # 1
BY analysis of FORALL,
WITH a FOR a:State
BY thinning (binding *)
BY simplification
BY synthesis of TRUE

```

```

--| Subproof # 2
  BY analysis of FORALL,
    WITH a FOR a:State
  BY thinning (binding *)
  BY simplification+
  BY synthesis of TRUE
--| end proof section
--| end Larch

```

5.2 More Complex Logic Circuit: Full Adder

A full adder represents the basic building logic structure to build more complex adder which adds two operands. Figure 5 shows the logic structure of the full adder.

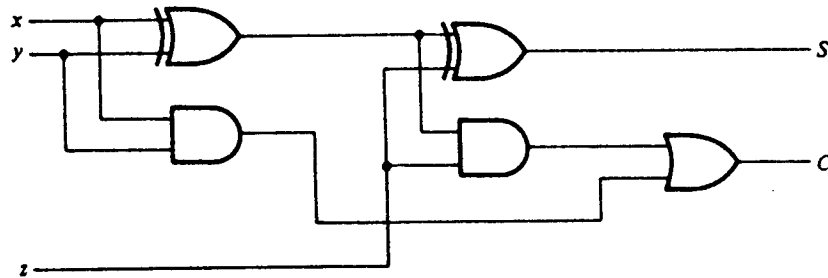


Figure 5: Full Adder

The proof structure for a full adder will be considered from two different approaches.

(1) The first approach is the architectural (logic gates) approach as shown.

```

entity full_adder is
  port(x, y, ci : bit; s, co : out bit);
end full_adder;

```

```

architecture arch of full_adder is
  component and2
    port(x,y : Bit; z : out Bit);
  end component;
  component ex_or
    port(x,y : Bit; z : out Bit);
  end component;
  component or2
    port(x,y : Bit; z : out Bit);
  end component;
  signal z1, z2, z3 : bit;
begin

```



```

11: ex_or port map(x,y,z1);
12: ex_or port map(ci,z1,s);
13: and2 port map(x,y,z2);
14: and2 port map(z1,ci,z3);
15: or2 port map(z2,z3,co);
end arch;

```

```

entity full_adder
guarantees
  always
    s = ( x xor y xor ci)
    with s delayed by 0 from x,y,ci
guarantees
  always
    co = ( ( x and y ) or (ci and (x xor y)))
    with co delayed by 0 from x,y,ci
end

```

The VCs for this logic structure is generated as shown.

```

--l Larch
proof section for arch_full_adder_VC
--l proof section
---> trait arch_full_adder_VC has unfinished proofs
--l vc in trait arch_full_adder_VC
--l proof:
>> (((Forall a:
  State::
    min(x'last_event @ a, y'last_event @ a) > 0
    -> z1 @ a = (x @ a xor y @ a)
    and z1'last_event @ a
    >= min(x'last_event @ a, y'last_event @ a) - 0)
  and (Forall a:
    State::
      min(ci'last_event @ a, z1'last_event @ a) > 0
      -> s @ a = (ci @ a xor z1 @ a)
      and s'last_event @ a
      >= min(ci'last_event @ a, z1'last_event @ a) - 0))
  and (Forall a:
    State::
      min(x'last_event @ a, y'last_event @ a) > 0
      -> (z2 @ a = '1') = ((x @ a and y @ a) = '1')
      and z2'last_event @ a
      >= min(x'last_event @ a, y'last_event @ a) - 0))
  and (Forall a:
    State::
      min(z1'last_event @ a, ci'last_event @ a) > 0
      -> (z3 @ a = '1') = ((z1 @ a and ci @ a) = '1')
      and z3'last_event @ a
      >= min(z1'last_event @ a, ci'last_event @ a) - 0))
  and (Forall a:
    State::
      min(z2'last_event @ a, z3'last_event @ a) > 0
      -> (co @ a = '1') = ((z2 @ a or z3 @ a) = '1')

```

```

        and co'last_event @ a
        >= min(z2'last_event @ a, z3'last_event @ a) - 0)
-> (Forall a:
  State::
    min(min(x'last_event @ a, y'last_event @ a), ci'last_event @ a)
    > 0
    -> s @ a = ((x @ a xor y @ a) xor ci @ a)
    and s'last_event @ a
    >= min(min(x'last_event @ a, y'last_event @ a),
      ci'last_event @ a) - 0)
and (Forall a:
  State::
    min(min(x'last_event @ a, y'last_event @ a), ci'last_event @ a)
    > 0
    -> co @ a
      = ((x @ a and y @ a) or (ci @ a and (x @ a xor y @ a)))
    and co'last_event @ a
    >= min(min(x'last_event @ a, y'last_event @ a),
      ci'last_event @ a) - 0)
<proof>
--| end proof section
--| end Larch

```

The complexity of the VCs is due to the complexity of the logical structure. It reflects all the logical components used to structure the full adder. The proof of this complex structure is straight forward and follows the same proof trend of the two-input AND logic gate shown previously.

```

--| proof:
  BY synthesis/analysis of FORALL/IMPLIES
  BY synthesis of AND
--| The first subproof
  BY synthesis/analysis of FORALL/IMPLIES
  BY thinning (binding *)
  BY simplification+
  BY synthesis of TRUE
--| The second subproof
  BY synthesis/analysis of FORALL/IMPLIES
  BY thinning (binding *)
  BY simplification+, rewriting+, simplification
  BY synthesis of TRUE
--| end proof section
--| end Larch

```

The proof structure of the full adder should be split into two parts using and synthesis as shown above. By using this command, the proof complexity is reduced to two subproofs due to the existence of the two Boolean functions: sum and carry generated in the architecture and in the specification. After that split, the proof structure follows the same trend used before to prove the subproof.

(2) The second approach is the behavioral approach as shown.

```

entity fadder is
  port (a, b, ci : in Bit; s, co : out Bit);

end fadder;

```

```

architecture behavior of fadder is
begin
  s <= (a xor b) xor ci;
  co <= (a and b) or ((a or b) and ci);
end behavior;

```

```

entity fadder
includes (Bit2Int)
  guarantees
  always
    int(a) + int(b) + int(ci) = 2*int(co) + int(s)
    with s, co delayed by 0 from a, b, ci
end

```

In this example, there is no logical gate in the architectural structure of the full adder. It is only the Boolean functions which generate the sum and the carry. The specification part indicates the general requirement of the full adder which is independent from the logical structure. That means, it is independent from the Boolean equations used to generate the sum and carry. It is clear that the specification of the behavior of the full adder needs the trait: Bit2Int to help in the proof structure and to make it more manageable to complete the proof. The VCs generated by Larch/VHDL theorem prover is shown below.

```

--l proof section
---> trait behavior_fadder_VC has unfinished proofs
--l vc in trait behavior_fadder_VC
--l proof:
>> (Forall a:
  State::
    ci'last_event @ a > 0
    and (b'last_event @ a > 0 and a'last_event @ a > 0)
    -> s @ a = ((a @ a xor b @ a) xor ci @ a)
    and s'last_event @ a
    >= min(ci'last_event @ a,
      min(b'last_event @ a, a'last_event @ a)))
and (Forall a:
  State::
    ci'last_event @ a > 0
    and (b'last_event @ a > 0 and a'last_event @ a > 0)
    -> co @ a
    = ((a @ a and b @ a) or ((a @ a or b @ a) and ci @ a))
    and co'last_event @ a
    >= min(ci'last_event @ a,

```

```

        min(b'last_event @ a, a'last_event @ a)))
-> (Forall a:
  State::
    (a'last_event @ a > 0 and b'last_event @ a > 0)
    and ci'last_event @ a > 0
    -> (int(a @ a) + int(b @ a) + int(ci @ a)
      = 2 * int(co @ a) + int(s @ a)
      and (s'last_event @ a
        >= min(min(a'last_event @ a, b'last_event @ a),
          ci'last_event @ a)
        and co'last_event @ a
          >= min(min(a'last_event @ a, b'last_event @ a),
            ci'last_event @ a)))
    <proof>
    --l end proof section
    --l end Larch

```

It is very clear that the VCs reflect the two major parts of the proof: the architecture behavioral part and the specification part. The proof structure is shown below.

```

--l proof:
  BY synthesis/analysis of FORALL/IMPLIES
  BY thinning (binding *)
  BY using fulladder in trait BitAdder
  rewriting left to right
  BY simplification+
  BY synthesis of TRUE
  --l end proof section
  --l end Larch

```

From the previous proof, it is clear that the proof structure follows the same trend used in the previous examples (logic gates). The only difference is the use of the theory of fulladder in the trait BitAdder defined in the specification which makes the proof easier to handle by providing the mathematical operations for converting Bit to Integer and vice versa.

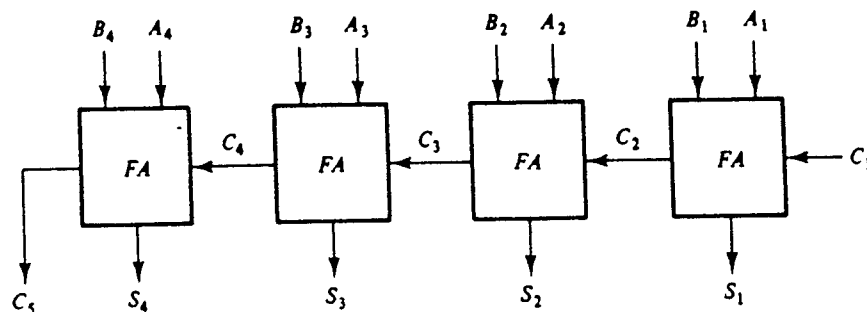


Figure 6: Four-Bit Ripple Adder

4.3 Complex Logic Structure: Four-Bit Ripple Adder

Figure 6 shows the logical structure of the four-bit ripple adder.

In this example, the logical complexity of the four-bit adder does not affect the proof structure as shown.

```
entity fourbit_adder is
  port(x1, y1, x2, y2, x3, y3, x4, y4, ci : bit; s1, s2, s3, s4, c1, c2, c3, co : out bit);
end fourbit_adder;
```

```
architecture arch of fourbit_adder is
  component full_adder
    port(x,y,ci : Bit; s,co : out Bit);
  end component;
  signal c1,c2,c3 : bit;
begin
  l1: full_adder port map(x1,y1,ci,s1,c1);
  l2: full_adder port map(x2,y2,c1,s2,c2);
  l3: full_adder port map(x3,y3,c2,s3,c3);
  l4: full_adder port map(x4,y4,c3,s4,co);
end arch;
```

```
entity fourbit_adder
guarantees
  always
    s1 = ( x1 xor y1 xor ci)
    with s1 delayed by 0 from x1,y1,ci
guarantees
  always
    c1 = ( ( x1 and y1 ) or (ci and (x1 xor y1))))
    with c1 delayed by 0 from x1,y1,ci
guarantees
  always
    s2 = ( x2 xor y2 xor c1)
    with s2 delayed by 0 from x2,y2,c1
guarantees
  always
    c2 = ( ( x2 and y2 ) or (c1 and (x2 xor y2))))
    with c2 delayed by 0 from x2,y2,c1
guarantees
  always
    s3 = ( x3 xor y3 xor c2)
    with s3 delayed by 0 from x3,y3,c2
guarantees
  always
    c3 = ( ( x3 and y3 ) or (c2 and (x3 xor y3))))
    with c3 delayed by 0 from x3,y3,c2
guarantees
  always
    s4 = ( x4 xor y4 xor c3)
    with s4 delayed by 0 from x4,y4,c3
guarantees
  always
    co = ( ( x4 and y4 ) or (c3 and (x4 xor y4))))
```

```

    with co delayed by 0 from x4,y4,c3
end

```

```

--| proof:
  BY synthesis of FORALL/IMPLIES
  BY synthesis of AND
    BY hypothesis
    BY hypothesis
    BY hypothesis
    BY hypothesis
    BY hypothesis
    BY hypothesis
    BY hypothesis
    BY hypothesis
  --| end proof section
--| end Larch

```

It is clear that the complexity of the logical structure shown previously does not prevent the proof to be very simple and straight forward. However, the complexity of the proof will be increased if the specification and the architecture use bit vector operation and iteration (for statement) in their construction as shown in Section 5.

4.4 Logic Structure with Time Delay

The following example of two-input AND gate shows the effect of signal delayed by an amount `del` on the proof structure. The amount of delay will be defined as a positive integer value as shown below.

```

entity and2_del is
  generic (del : positive);
  port ( x, y : Bit; z : out Bit);
end and2_del;

```

```

architecture arch of and2_del is
begin
  z <= x and y after del;
end arch;

```

```

entity and2_del
  guarantees
  always
  z = (x and y)
  with z delayed by del from x, y
end

```

```

--> trait arch_and2_del_VC has unfinished proofs
--| Larch
proof section for arch_and2_del_VC
--| proof section
--> trait arch_and2_del_VC has unfinished proofs
--| vc in trait arch_and2_del_VC

```

```

--| proof:
  >> (Forall a:
    State::
      min(y'last_event @ a, x'last_event @ a) > del
      -> z @ a = (x @ a and y @ a)
      and z'last_event @ a
      >= min(y'last_event @ a, x'last_event @ a) - del)
  -> (Forall a:
    State::
      min(x'last_event @ a, y'last_event @ a) > del
      -> z @ a = (x @ a and y @ a)
      and z'last_event @ a
      >= min(x'last_event @ a, y'last_event @ a) - del)
  <proof>
--| end proof section
--| end Larch

```

```

--| proof:
  BY synthesis/analysis of FORALL/IMPLIES
  BY thinning (binding *)
  BY simplification+
  BY synthesis of TRUE
--| end proof section
--| end Larch

```

Comparing the two previous proofs (and2 and and2_del), it is clear that there is no difference in the proof structure of the two-input AND gate with/without delay. The difference appears when the delay is used with a more complex logic circuit as shown in the following example of the sum output of a full adder with delay as shown below.

```

entity sum_del is
  generic ( del : positive );
  port(x, y, ci :bit; s : out bit);
end sum_del ;

```

```

architecture arch of sum_del is
  component xor_del
    generic ( del : positive );
    port(x,y : Bit; z : out Bit);
  end component;
  signal z1 :bit;
  begin
    l1: xor_del generic map (del) port map(x,y,z1);
    l2: xor_del generic map (del) port map(ci,z1,s);
  end arch;

```

```

entity sum_del
  guarantees
    always
      s = ( x xor y xor ci)
      with s delayed by 2*del from x,y,ci

```

end

```
---> trait arch_sum_del_VC has unfinished proofs
--| vc in trait arch_sum_del_VC
--| proof:
  >> (Forall a:
    State::
      min(x'last_event @ a, y'last_event @ a) > del
      -> z1 @ a = (x @ a xor y @ a)
      and z1'last_event @ a
      >= min(x'last_event @ a, y'last_event @ a) - del)
    and (Forall a:
      State::
        min(ci'last_event @ a, z1'last_event @ a) > del
        -> s @ a = (ci @ a xor z1 @ a)
        and s'last_event @ a
        >= min(ci'last_event @ a, z1'last_event @ a) - del)
  -> (Forall a:
    State::
      min(min(x'last_event @ a, y'last_event @ a), ci'last_event @ a)
      > 2 * del
      -> s @ a = ((x @ a xor y @ a) xor ci @ a)
      and s'last_event @ a
      >= min(min(x'last_event @ a, y'last_event @ a),
        ci'last_event @ a) - 2 * del)
    <proof>
  --| end proof section
--| end Larch
```

```
--| proof:
  BY synthesis/analysis of FORALL/IMPLIES
  BY thinning (binding *)
  BY using del_constraint as new hypothesis
  BY simplification+
  BY synthesis of TRUE
--| end proof section
--| end Larch
```

It is clear from this example that the amount of delay `del` is defined as two `xor_del` units delay. The proof structure is similar to the previous one for the two-input AND gate with delay. The only difference is the command: `BY using del_constraint as new hypothesis` which introduces the delay as a new hypothesis to be used by the simplifier after this command. Without this new hypothesis, which has been built into the library, the simplifier will not be able to simplify the proof to `TRUE`.

5. The Algorithmic Nature of the Proof

From the above examples, any logic circuit structured using logical gates has similar proof steps. It is very interesting to notice the following rules when constructing the proof for a logic circuit.

- (1) The proof for any logical entity should be started by the command: **By synthesis of FORALL/IMPLIES**.
- (2) The above step can be simplified more by adding the command: **with analysis** to the previous synthesis command.
- (3) The proof may be simplified directly using the command: **SDVS simplification with add +**, which could end the proof correctly with: **By synthesis of TRUE**. However, if the logical entity is very complex, then another step must be taken using the command: **By synthesis of AND**, which will split the proof into several subproofs as shown in the above examples. In this case, each subproof must be proved separately using the same proof methodology explained above.
- (4) If the logic circuit has a delay which is defined as `del` , then the command **BY** using `del_constraint` as new hypothesis, which must be used before the simplification process. Also, it is very important to count correctly the number of units delay which must be inserted in the specification as shown. If the number of delay counts was less than the exact number of units delay, the prover will not be able to prove the correctness of the specification.

It is importance to notice that in some cases the command: **analysis of FORALL** may be omitted from the proof steps shown previously.

6. Special Cases: Complex Proof Structure

The proof start to deviate from the previous algorithmic method when the entity and/or its architecture contains vectors and/or iteration using for statement as shown in the following examples.

6.1 Example of A Vector Signal Assignment

This second example is a simple modified version of the a simple Bit-assignment; it is a Bit-vector assignment. The proof constructed by ORA Larch/VHDL became complex and not easy to be followed in a systematic way as shown below.

```
entity vector1 is
  generic (n : positive);
  port ( x : Bit_vector ((n-1) downto 0) ; z : out Bit_vector ((n-1) downto 0));
end vector1 ;
```

```

architecture arch of vector1 is
begin
    z (n-1 downto 0) <= x (n-1 downto 0);
end arch;

```

```

entity vector1
includes(StableArray)
guarantees
    always
        z = x
        with z delayed by 0 from x
end

```

Verification Conditions

```

>> (Forall a:
State::
    slice(x, n - 1, 0)'last_event @ a > 0
    -> slice(z @ a, n - 1, 0) = slice(x @ a, n - 1, 0)
        and slice(z, n - 1, 0)'last_event @ a
            >= slice(x, n - 1, 0)'last_event @ a - 0)
-> (Forall a:
State::
    x'last_event @ a > 0
    -> z @ a = x @ a and z'last_event @ a >= x'last_event @ a - 0)

```

--| **Proof:**

```

BY synthesis/analysis of FORALL/IMPLIES
BY analysis of IMPLIES.
    BY analysis of FORALL.
        WITH a FOR a : State
    BY analysis of IMPLIES.
        BY using last_event_slice in trait Bit_StableArray with (x for x,
a for a, n-1 for i, 0 for j) as new hypothesis
        BY simplification
        BY synthesis of TRUE
    AND THEN
        BY hypothesis
    AND THEN
        BY synthesis of AND
        BY thinning (not 3)
        BY using eq in trait Vector with (z @ a for a,
x @ a for b) as new hypothesis
        BY rewriting
        BY claiming Forall i : Int::
in_domain(i, z @ init)->(z @ a)[i] = (x @ a)[i]
            BY synthesis/analysis of FORALL/IMPLIES
            BY thinning 2
            BY using in_domain as rewrite rule
            BY using valid_slice as rewrite rule
            BY using low in trait Vector as rewrite rule
            BY using high in trait Vector as rewrite rule
            BY using slice_val in trait Vector with (z @ a for a,
n-1 for i, 0 for j, i for n) as new hypothesis
            BY using slice_val in trait Vector with (x @ a for a,

```

n-1 for i, 0 for j, i for n) as new hypothesis
 BY rewriting (6), simplification
 BY synthesis of TRUE
 THEN
 BY simplification
 BY synthesis of TRUE

 BY thinning (not 4)
 BY claiming slice(z, n-1, 0)'last_event @ a = z'last_event @ a
 BY using last_event_slice in trait Bit_StableArray with (z for x,
 a for a, n-1 for i, 0 for j) as new hypothesis
 BY claiming slice(z, n-1, 0)'last_event @ a <= z'last_event @ a
 BY using last_event_components
 in trait Bit_StableArray with (z for x, a for a,
 slice(z, n-1, 0)'last_event @ a for n) as new hypothesis
 BY claiming slice(z, n-1, 0)'last_event @ a <= time(a)
 BY rewriting
 BY synthesis of TRUE
 THEN
 BY claiming Forall i : Int::
 in_domain(i, z @ a)
 -> z[i]'last_event @ a >= slice(z, n-1, 0)'last_event @ a
 BY synthesis/analysis of FORALL/IMPLIES
 BY thinning mentioning time
 BY using last_event_slice_component
 in trait Bit_StableArray with (z for x, a for a, n-1 for i, 0 for j,
 i for k) as new hypothesis
 BY analysis of IMPLIES,
 BY using in_domain as rewrite rule
 BY using valid_slice as rewrite rule
 BY using low in trait Vector as rewrite rule
 BY using high in trait Vector as rewrite rule
 BY rewriting (6), simplification
 BY synthesis of TRUE
 AND THEN
 BY using last_event_component
 in trait Bit_StableArray with (slice(z, n-1, 0) for x, a for a,
 i for i) as new hypothesis
 BY simplification
 BY synthesis of TRUE
 THEN
 BY simplification
 BY synthesis of TRUE
 THEN
 BY simplification
 BY synthesis of TRUE
 THEN
 BY claiming slice(x, n-1, 0)'last_event @ a = x'last_event @ a
 BY using last_event_slice
 in trait Bit_StableArray with (x for x, a for a, n-1 for i,
 0 for j) as new hypothesis
 BY claiming slice(x, n-1, 0)'last_event @ a
 <= x'last_event @ a
 BY using last_event_components
 in trait Bit_StableArray with (x for x, a for a,

```

        slice(x, n-1, 0)'last_event @ a for n) as new hypothesis
    BY claiming slice(x, n-1, 0)'last_event @ a<=time(a)
    BY rewriting
    BY synthesis of TRUE
    THEN
    BY claiming Forall i : Int:
        in_domain(i, x @ a)
        -> x[i]'last_event @ a>=slice(x, n-1, 0)'last_event @ a
    BY synthesis/analysis of FORALL/IMPLIES
    BY thinning mentioning time
    BY using last_event_slice_component
        in trait Bit_StableArray with (x for x, a for a, n-1 for i, 0 for j,
        i for k) as new hypothesis
    BY analysis of IMPLIES,
    BY using in_domain as rewrite rule
    BY using valid_slice as rewrite rule
    BY using low in trait Vector as rewrite rule
    BY using high in trait Vector as rewrite rule
    BY rewriting (6), simplification
    BY synthesis of TRUE
    AND THEN
    BY using last_event_component
        in trait Bit_StableArray with (slice(x, n-1, 0) for x, a for a,
        i for i) as new hypothesis
    BY simplification
    BY synthesis of TRUE
    THEN
    BY simplification
    BY synthesis of TRUE
    THEN
    BY simplification
    BY synthesis of TRUE
    THEN
    BY simplification
    BY synthesis of TRUE

```

The proof presented previously is very complex compared with the proof of the previous examples. The question is: why? It is very difficult to answer this question and find algorithmic way to proof similar entity. The only method to be used in such a case is to formulate a theory for the above proof so that this theory becomes general enough to be applied to similar cases when a Bit-vector is used. The following example is more complex and show the iterative form of a four-bit Adder using for loop to repeat the structure of the four full adders.

6.2 Example of Iterative Four-Bit Ripple Adder Using For Loop

```

entity add_4 is
    port ( x, y : Bit_vector (3 downto 0); ci : Bit_vector (4 downto 0);
          s, co : out Bit_vector (3 downto 0));
end add_4;

```

```

architecture arch of add_4 is
component full_adder
    port(x,y,ci : Bit; s,co : out Bit);
end component;
begin
    d0to3 : for i in 0 to 3 generate
        d0 : full_adder port map(x(i),y(i),ci(i),s(i),co(i));
        ci(i+1) <= co(i);
    end generate;
end arch;

```

```

entity add_4
includes ( BitVector)
guarantees always
    forall i : Int :: 0 <= i and i <= 3 ->
        s[i] = (x[i] xor y[i] xor ci[i])
        with s delayed by 0 from x, y, ci
guarantees always
    forall i : Int :: 0 <= i and i <= 3 ->
        co[i] = ( (x[i] and y[i]) or (ci[i] and (x[i] xor y[i])))
        with co delayed by 0 from x, y, ci
guarantees always
    forall i : Int :: 0 <= i and i <= 3 ->
        ci[i+1] = co[i]
        with ci delayed by 0 from co
end

```

The proof of the above iterative four-bit adder structure is more complex than the proof of the simple vector assignment statement shown in the previous Example. It is also possible to make the specification of four-bit adder more general as shown below.

```

entity add4
includes(Int4)
guarantees
    always
        int4bit(s[3],s[2],s[1],s[0]) + int(co)*2**4 = int4bit(x[3],x[2],x[1],x[0]) + int4bit(y[3],y[2],y[1],y[0])
        with s[3],s[2],s[1],s[0], co delayed by 0 from x[0],x[1],x[2],x[3],y[0],y[1],y[2],y[3]
end

```

The specification of a four-bit adder independent of the architecture (logical circuit design) which make the proof more complex than before. Those special and more general cases needs more research work to be done to reduce their complexity [13-15].

7. Formal Methods in Education

The formal verification techniques are composed at a very high level of abstraction and they have never been used on a wide scale to verify hardware and software design projects. Very intensive research projects on

formal methods have been sponsored and funded by the DARPA Information Technology Office (ITO), the Air Force Research Facility at Rome Laboratory, and NASA. Also, Rome Laboratory and DARPA's ITO have established the 21st Century Engineering Consortium. The reason for establishing the consortium is to promote teaching of formal methods at all academic levels (undergraduate and graduate) so that the engineering capability necessary to build highly assured systems can be developed. This means that DARPA and the Air Force would like to see a much larger number of students trained nationwide in engineering and computer science using formal methods (examples: hardware verification using Larch/VHDL, software verification using ADA, C, C++ codes, requirements/specification analysis, as well as applying this technology to other problem domains). The following topics are the main themes of the Consortium:

- (1) Who will be the customer for our formal method products?
- (2) How will we educate students in formal methods?
- (3) What are we going to teach?
- (4) How will we achieve our goals?
- (5) What formal method materials must be taught in the undergraduate and graduate level courses?
- (6) How can formal methods be integrated into existing MATH/CS/CE courses?

A model course in Formal Specification and Verification with Laboratory at the undergraduate level using theorem provers has been suggested as shown in [16,17]. This course offers a new approach for students and teachers alike in which theorem provers, like Larch/VHDL, will be used to help students structure and understand the proof methodology. A training session which will focus on how to use the technology (software packages selected) and how to formulate problems (writing correct specifications) so that the theorem prover can proceed to proof its correctness was also suggested. This teaching methodology suggested in [16,17] will encourage faculty members to introduce formal verification and proof methodology into their courses and then a wide range of undergraduate students will be trained and equipped to deal with more complex verification provers in the future. The 21st Century Engineering Consortium announced its first workshop which will be held in Melbourne, Florida, March 18-19, 1998. This workshop's principal concern is promoting formal methods in computer science and engineering program and integrating it in the main stream curricula of these disciplines.

8. Conclusion

The most important feature of the Theorem Prover developed by ORA is its capability to guide the user to structure the proof to the problem. So Larch/VHDL theorem prover is a very important educational tool to teach the mathematical foundation to the student and encourage them to proof any problem if it is possible to write its specification correctly. That means, if the problem is formulated and its specification is written in the correct way. The proof structure of different classes of logical circuits was developed in an algorithmic way. The complexity problem of the proof of certain logic structure was addressed and solution was suggested to improve the performance of the theorem prover. By providing more proved theories into the library body of the existing theorem prover, the complexity of the proof can be reduced to its minimum. It was expected that formal method for specification and verification of both hardware and software projects will overcome its difficulties and dominate the field of producing reliable and secure systems.

9. References

- [1] IEEE Standard VHDL Language Reference Manual. ANSI/IEEE Std 1076-1993. June 4, 1994.
- [2] IEEE Standard for Waveform and Vector Exchange (WAVES). IEEE Std 1029.1-1992, and IEEE Standard Multivalued Logic System for VHDL Interoperability (std_logic_1164), May 26, 1993.
- [3] Z. Navabi, VHDL: Analysis and Modeling of Digital Systems, McGraw-Hill, 1993.
- [4] P. J. Ashenden, The Designer's Guide to VHDL, Morgan Kaufmann, 1996.
- [5] D. Jamsek and M. Bickford, "Formal Verification of VHDL Models," Final Technical Report, Rome Laboratory RL-TR-94-3, March 1994.
- [6] Odyssey Research Associates, "Penelope Reference Manual V3-3," TM94-0009, December 1993.
- [7] M. Bickford, "Technical Information Report: Final Report for Formal Verification of VHDL Design," Odyssey Research Associates, F30602-94-C-0136, CDRL, A005, Rome Laboratory/ERDD, July 1996.
- [8] M. Bickford, "Technical Information Report: User/Training Manual for Formal Verification of VHDL Design," Odyssey Research Associates, F30602-94-C-0136, CDRL, A004, Rome Laboratory/ERDD, 621 July 1996.

- [9] S. Garland, J. Guttag and J. Horning, "Debugging Larch Shared Language Specification," IEEE Trans. on Software Engineering, Vol. 16, no. 9, September 1990.
- [10] J. Guttag, J. Horning and J. Wing, "The Larch Family of Specification Languages," IEEE Software, September 1985.
- [11] J. Wing, "Writing Larch Interface Language Specifications," ACM Trans. on Programming Languages and Systems, Vol. 9, no.1, January 1987.
- [12] J. V. Guttag and J. J. Horning, LARCH: Languages and Tools for Formal Specification, Springer-Verlag, 1993.
- [13] A. E. Barbour, "Formal Verification Using ORA Larch/VHDL Theorem Prover," final report for Summer Faculty Research Program, Air Force Office of Scientific Research and Rome Laboratory/ERDD, August 1996.
- [14] A. E. Barbour and M. P. Nassif, "Basic Concepts of Hardware Formal Verification Using ORA Larch/VHDL," Proceedings of the International Conference on Parallel and Distributed Processing Techniques and Applications, Vol. II, pp. 778-782, Las Vegas, Nevada, June 30 - July 3, 1997.
- [15] A. E. Barbour and M. P. Nassif, "Hardware Formal Verification Using ORA Larch/VHDL Theorem Prover," presented in the 1997 Summer Computer Simulation Conference, Key Bridge Marriott, Arlington, Virginia, July 13-17, 1997.
- [16] A. E. Barbour and L. J. Olszewski, "An Introductory Course in Formal Specification and Verification With A Computer Lab," Proceedings of the Eleventh Annual Consortium for Computing in Small Colleges Southeastern Conference, pp. 88-99, Lenoir-Rhyne College, Hickory, NC, November 7-8, 1997.
- [17] A. E. Barbour and L. J. Olszewski, "Education in Formal Methods Using Theorem Provers," sent for publication at the Second KFUPM Workshop on Information & Computer Science - WICS'98, March 10, 1998.
- [18] M. Bickford and D. Jamsek, "Larch/VHDL Familiarization," Course Notes presented at Syracuse University, NY, May 12-14, 1997.

FORMAL SPECIFICATION AND VERIFICATION
OF
MISSI RECEIVER AND FORWARDING USING SPIN

Milica Barjaktarović
Assistant Professor
Department of Electrical and Computer Engineering

Wilkes University
Stark Learning Center
Wilkes-Barre PA 18766

Final Report for:
Summer Research Extension Program

Sponsored by:
Air Force Office of Scientific Research
Bolling AFB

and
Wilkes University

June, 1997

FORMAL SPECIFICATION AND VERIFICATION OF MISSI RECEIVER AND FORWARDING USING SPIN

Milica Barjaktarović
Assistant Professor
Department of Electrical and Computer Engineering
Wilkes University
Wilkes-Barre PA 18766

Abstract

In this document we formally specify and verify a part of the Multilevel Information System Security Initiative (MISSI). MISSI is a National Security Agency (NSA) program, designed to send protected messages over unprotected networks such as Internet. MISSI uses several kinds of cryptography for protecting the messages. Cryptography is accomplished using a credit card sized Personal Computer Memory card Interface Association (PCMCIA) card called the FORTEZZA Crypto Card (or "Card", for short).

We constructed a formal specification of receiving and forwarding e-mail using MISSI. We used ACP 123, X.400 and SDN 701 standards to specify message structure and MISSI access and forwarding policy. We used formal language called Promela, based on Hoare's CSP. We verified the model using automated model checker SPIN, developed by AT&T.

FORMAL SPECIFICATION AND VERIFICATION OF MISSI RECEIVER AND FORWARDING USING SPIN

Milica Barjaktarović

1 Introduction

The Multilevel Information System Security Initiative (MISSI) is a National Security Agency (NSA) program, designed to send protected messages over unprotected networks such as Internet [2]. New releases of MISSI are posted in

<http://www.armadillo.huntsville.al.us>.

MISSI employs cryptography by using a credit card sized Personal Computer Memory card Interface Association (PCMCIA) card called the FORTEZZA Crypto Card (or "Card," for short). MISSI Forteza cards handle Sensitive but Unclassified (SBU) messages, and FORTEZZA Plus Cards handle Secret (S) messages. MISSI uses a combination of private and public cryptography. Users' public cryptography information is posted in the public distributed database called the Directory, in files called certificates [5, 12].

MISSI messages can travel using X.400 or MIME mail. MISSI messages which travel using X.400 must be sent as military messages (MMs). MMs can be individual or organizational. Organizational messages commit organizations, and require special security concerns. For example, if A sends an organizational military message to B, can B forward it to a non-organizational user?

From the engineering perspective, the two basic questions are:

- how do we design and build MISSI receiver and forwarding policy?
- how do we know that it works correctly?

In this report, we will discuss the above questions. This report is a continuation of our work presented in [6] and [7].

The basis of our work is application of a mathematically based formal method to specify and verify local cache management protocol, in order to gain better understanding of protocol design by formal reasoning and proofs.

The goals of our work are:

- help the designers of complex systems such as MISSI comprehend and design the system in an organized way
- help the communication between teams designing and building various parts of the system

- eventually hand the verified formal specification to the programmers and builders; based on their input, change the specification, and repeat this step.

We chose formal language Promela, based on Hoare's CSP, to specify and test MISSI receiving and forwarding. Promela is a language based on CSP and implemented in C. SPIN is the publicly distributed automated verification tool which supports Promela [8, 10, 9, 15].

We have cited various documents throughout this report, and included our comments in italics.

2 Overview of MISSI

MISSI provides the following security services: data integrity and authentication (via hashing, digital signatures, and certificate authentication); confidentiality (via public and secret cryptography); non-repudiation with proof of origin; and non-repudiation with proof of receipt (optional). Users are authorized by certification authorities, which issue certificates ("tickets") to the users.

In order to illustrate how MISSI works, we will assume that User A wants to send e-mail to User B. A will login with his Card, supply his PIN number, write a text message, and his Card will encrypt the message using a secret key and sign the message. The secret key is encrypted using public key cryptography and sent in the header of the message. Receiver will extract the secret key using public decryption, and decrypt the message. Receiver will also verify sender's signature, and signatures of authorities that signed sender's signature.

The Card is contacted via SDN.701 Message Security protocol. The process of sending out a message to the Internet involves using X.400 or RFC1521 message formatting mechanisms. X.400 is a suite of standards for message formatting and transfer. Each workstation has application software called User Agent (UA), which sends and receives messages.

Public keys are posted in the MISSI Directory. Each user's private keys and PIN are stored on the user's Card and non-readable by the user. X.500 is a suite of standards for the distributed Directory service. X.500 Directory is a distributed "white/yellow pages" repository of public keys. The MISSI Directory is the X.500 Directory equipped with a FORTEZZA card. UA contains Directory User Agent (DUA), which contacts the Directory through Directory Service Agent (DSA).

Public keys are stored in **certificates**. A certificate is a data structure which includes each user's name and public keys, signed by the authority that issued the certificate. Certificates and keys can expire, get compromised in some way, or be revoked. Authorities keep invalid certificates' serial numbers in Certificate Revocation Lists (CRLs). Authorities also keep Key Revocation Lists (KRLs). X.509 standard specifies the certification hierarchy and certificate and key management.

On top of the hierarchy is the Policy Approving Authority (PAA), which is the trusted authority. PAA's signature is stored on each user's Card, and used for the final authentication. PAA issues certificates for Policy Creation Authorities (PCAs), but cannot revoke any PCA and does not keep

Certificate Revocation List (CRL) for PCAs. Each PCA issues certificates for Certification Authorities (CAs) in its domain, and maintains a CA CRL. It also maintains a CA KRL, and distributes it to all CAs in its domain. PCA will post signed CRL and KRL to the Directory. Each CA issues user certificates, and maintains and posts user CRL to the Directory.

PAA, PCA and CA operate through a CA workstation (CAW) which can be set up to provide PAA, PCA or CA capabilities. CAW has a trusted operating system, and is connected to a FORTEZA device.

In Figure 1 we show the overall MISSI architecture and identify MISSI components. The figure represents an enclave consisting of a Local Area Network (LAN) with MISSI components, attached to the Internet via a Secure Network Server (SNS). SNS ensures that messages of appropriate security levels leave and enter the enclave. We are assuming the enclave is classified as Secret. Sensitive but Unclassified enclaves can have a commercial firewall instead of SNS [13]. Other enclaves are attached to the WAN, but we do not show them to conserve the space.

Workstations are equipped with either FORTEZZA Cards (F) or FORTEZZA Plus cards (FP), or they have no cards. Users are certified by the Certification Authority (CA) operating at the CA workstation (CAW). Audit Manager (AM) is used for auditing the system. Mail List Agent (MLA) is used to forward mail to a list of recipients. Rekey Manager (RKM) is used to rekey the cards.

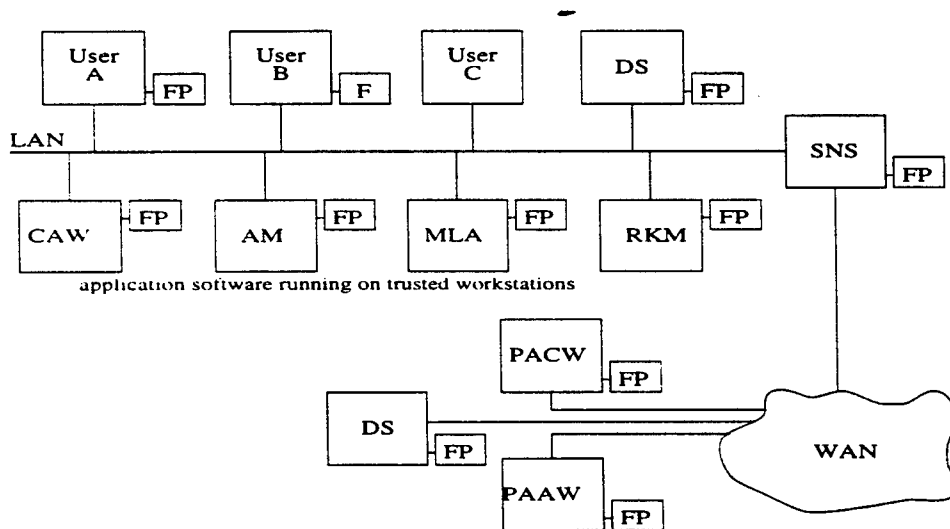


Figure 1: MISSI Components

3 MISSI Workstation

The abstract view of the interface between a MISSI workstation and FORTEZZA card consists of three interacting units: UA, MSP functions, and the Card. UA application process contacts the

hardware on the Card through MSP application software. MSP application software represents a high-level interface to the Card and consists of a library of eight `msp_` functions. These functions call on the library of fifty one Crypto Interface (CI) `CI_` functions, which interface with the card device driver and eventually with the hardware. The interfaces are shown in Fig. reffig:3-rings.

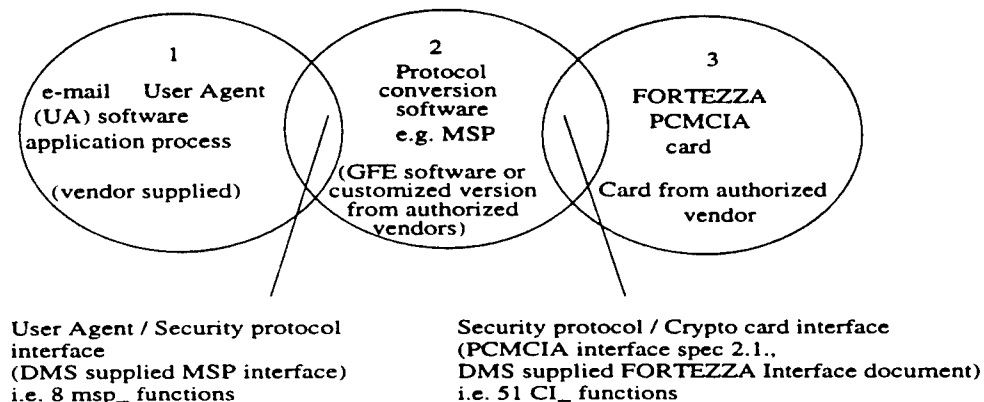


Figure 2: Abstract View of MISSI Workstation /FORTEZZA Card Interface

In Figure 3 we show the software components within a MISSI workstation and its connection to the FORTEZZA hardware.

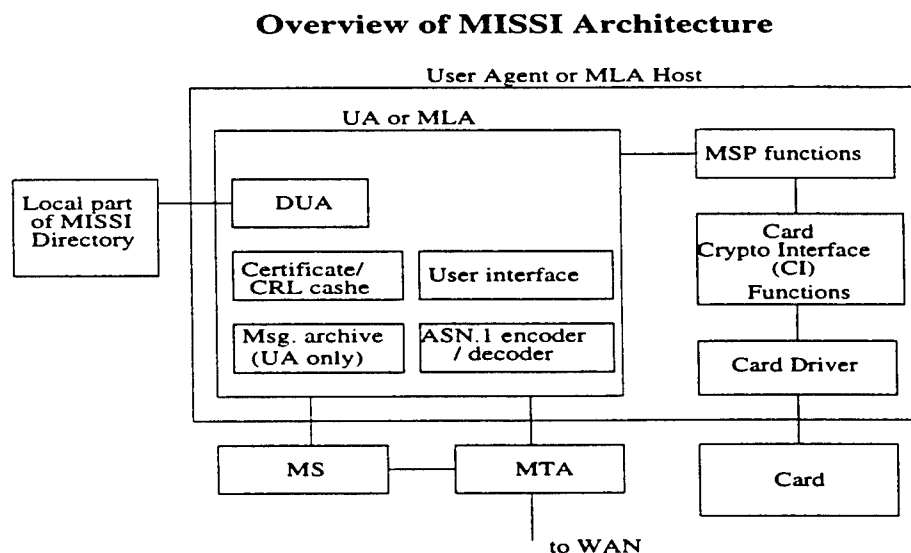


Figure 3: Local Workstation Components

Certificate validation is described in detail in [7].

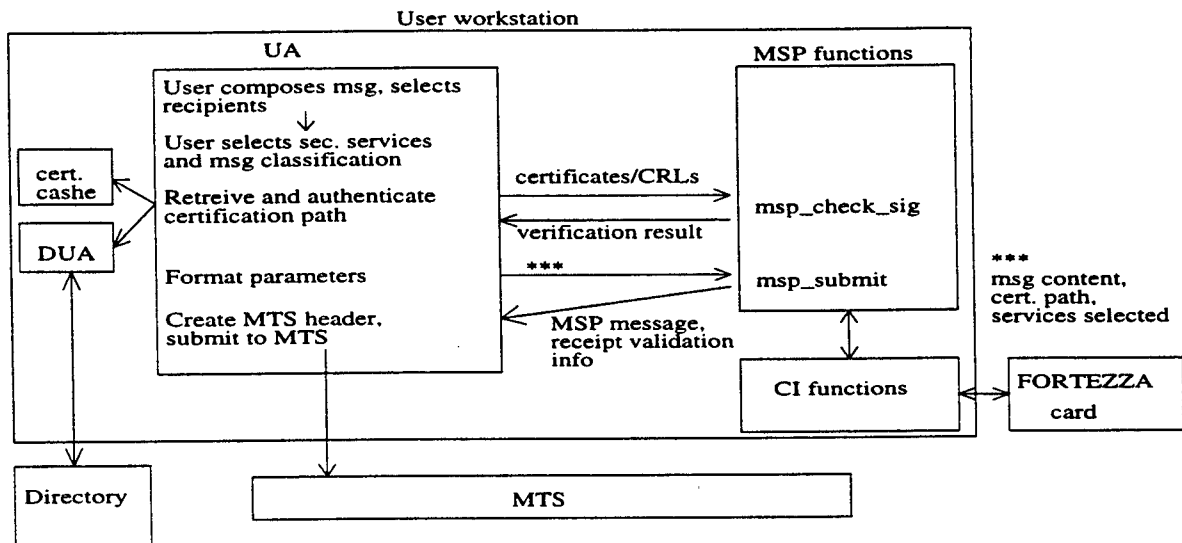


Figure 4: MISSI Send Mail

4 Message Headers

If MISSI uses X.400 for message transport, all messages must be formatted as Military Messages (MMs), which are labeled with content type P772. The nesting of X.400, MSP and P772 messages is shown in Fig. 5.

This section outlines the sections of RFC822/1521 and MSP headers which are crucial for processing of received and/or forwarded messages, in the most general terms. We do not include all header fields, for the sake of brevity.

RFC822, or RFC 1521 message contains the following fields: RFC822/1521 header, and body part (which is ANS.1 encoded MSP message).

RFC822/1521 header information is taken from the summary provided in [3], and the official MSP header specification is taken from [5].

RFC822/ 1521 header contains the following fields:

From: sender DN

To: receiver DN

Services applied (optional)

Content type: multipart/X-MSP

Preamble: seems like it does not contain useful MISSI info

Plain text: user-selectable option, if message not encrypted

Body part:

application type: MSP

security applied: signed /signed, encrypted

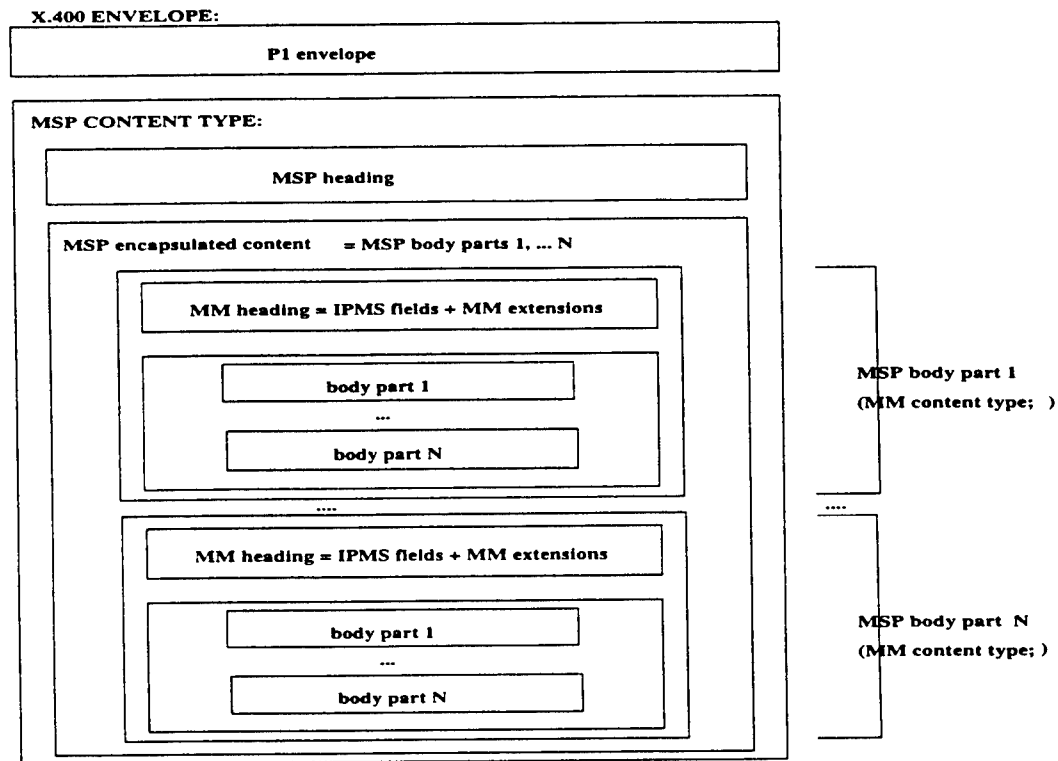


Figure 5: Message Nesting

ANS.1 encoded MSP message consists of MSP header and MSP encapsulated content. It is possible that the header and the encapsulated content are signed (this signature is called sequence signature, and the message is called a signed MSP message). `msp_status` and `msp_deliver` process this “MSP” part of the incoming message.

MSP header contains the following fields: **MessageSecurityProtocol**: Msp / SIGNED Msp

“Msp” field is a sequence of the following fields:

OriginatorSecurityData (optional): contains information needed to process `PerRecipientToken`, i.e. for decryption. This information will be either pertaining to the original sender, or to MLA if MLA processed the message.

....

originatorCertificate (optional): this field is required according to FORTEZZA requirements [4].

organizational msg flag is set in the certificate

additionalSecurityInfo (optional): contains `lrbac` information

....

signatureBlock (optional): originator’s signature information, in case the message was signed

recipientSecurityData: set of PerRecipientToken (optional):

....

RecipientKeyToken:

....

encapsulatedContentType:

SecurityLabel:

....:

SecurityClassification (optional): U/ UBS/ C/ S/ TS

SecurityCategory (optional): contains prbac

....

contentDescription

mspSequenceSignatureCertificate (optional): certification path of the originator

encapsulatedContent

What is missing here is the information about the security level of the originator's enclave. Some MISSI decisions are based on the level of the originator's and recipient's enclaves, yet this information is not carried in MSP header. Also, security classification is an optional field in SDN.701, yet some MISSI decisions are based on it. Therefore, we conclude that MISSI messages must have security classification specified.

[11], p.14, specifies that "UAs are grouped into classes based on the type of content of messages they can handle. The MTS provides a UA with the ability to identify its class when sending messages to other UAs."

5 ACP 123: Military Messaging

Military messaging is used in Defense Messaging System (DMS) [16], [17]. All Military Messages (MMs) must have security service applied to them.

ACP 123 defines extensions to IP messaging which supports Military messaging. "Unless exceptions are noted, all statements which apply to Interpersonal Messaging Services (IPMS) also apply to Military Messaging Services (MMS)" [17], p.A-2. Military messaging is a superset of interpersonal messaging. "The structure of military messages is fundamentally the same as that of IP messages. Additional elements in the heading are required to support MM, as shown in Fig. 5. IP messages (IPMs) have content type P22, and military messages (MMs) have content type P772" [17], p.A-5.

MM-UA must keep a record for each message. For incoming messages, the following record is kept:

- security label
- extended authorization information

- ...
- message originator

For outgoing messages, the following record is kept:

- security label
- extended authorization information
- ...
- release authority

It is not clear from ACP 123 standard and its Annex A if MMs must be organizational, or individual messages are allowed as well. Seems like ACP 123 allows only organization messages, and Annex A allows individual messages as well:

“ACP 123 identifies the service and protocol requirements necessary to ensure interoperability among Military Message Handling Systems (MMHS) for military message traffic that formally commits an organization and requires authorized release. This ACP specifically does not cover messaging between individuals” [16], p.1-1.

However, Annex A of ACP 123, aka US Supplement, does allow individual messaging: “The Military Messaging Service (MMS) provides an electronic mail facility between military personnel and military organizational users. ...

The Military Messaging User Agent (MM-UA) is the entity that represents a military individual or organization with the MMS.

The MMS protocol supports the exchange of messages between military individuals or organizations. The nomenclature for this protocol is P772” [17], p.A-9.

The question is: are military personnel who are not organizational users a special kind of individual users?

5.1 Individual and Organizational Users

There are two classes of users in the DMS: individual and organizational. These two kinds of users must be distinguished, as security requirements are greater for organizational users. The organizational or individual classification is assigned through digital signature privileges in the Digital Signature Standard (DSS) key material.

Organizational users represent an organization and can commit organizational resources. Organizational messaging is defined as message exchange between organizational elements. These messages require approval for transmission by designated officials at the sender side, and determination of distribution by the receiver side. [5] requires all MSP UAs to “determine whether a message is an organizational message and unambiguously communicate that fact to the user. A message is

considered an organizational message iff it has been signed by a user with a certificate designated as an organizational certificate." Therefore, we conclude that *organizational users must not send unprotected messages*; and that *organizational users must be at least SBU MISSI users*.

All other messages are classified as individual messages.

A user's certificate contains the following fields of interest to us:

KEA Clearance field, containing user security clearance (U, SBU, S, or TS);

DSS Privileges field, containing user's privileges (organizational authority, PAA, PCA, PA, or NoSignatureCapability/ReadOnly).

6 Receiver Requirements

[4], p.12, specifies:

"12. Upon successful verification of an MSP protected message, the application shall:

- a. Display to the recipient the originator's authenticated DN.
- b. Provide the recipient with the option to display the entire certificate hierarchy in a signature notification report. *Which certification hierarchy - there can be three of them included within MSP header: sender's, SNSs (for sequence signature), and MLAs (for encryption token).*
- c. Display to the recipient the MSP security label of an encrypted message after successful message verification.

The application shall check the MSP security label to verify that both the originator and recipient have that clearance authorized in their certificates before displaying the message.

If the security label check fails, the user shall be notified that the message clearance is unauthorized, message processing shall stop and any message processing data shall be deleted and overwritten.

- d. Identify the message as organizational to the recipient if the Organizational Releaser privilege bit is set in the originator's certificate.
- e. Display to the recipient only the message header and content that was MSP protected. "

"15. Perform MSP sequence signature validation prior to any other MSP processing when receiving a message containing a sequence signature.

The application shall notify the user if the originator's DN for the sequence signature is not the same as the originator's authenticated DN for the message signature.

If the authenticated DNs are different then the application shall provide the recipient the option to display both DNs and certificate hierarchies."

"16. Include the entire user to PCA certificate path in MSP header by default." *Since there can be up to three certification paths in a message, should we include them all or only the sender's path?*

[1] p.42 states that: "To comply with SDN.701, the Application (i.e. UA) must unambiguously communicate the following information to the user:

1. The authenticated identity of the message signer, if the message was signed
2. The authenticated identity of the message encryptor (i.e. token generator)
this could be either the sender or MLA - or SNS?, and the security label
 if the message, if the message was encrypted
3. any errors detected during MSP processing.”

7 Receiver Algorithm

MISSI receiver is shown in Fig. 6.

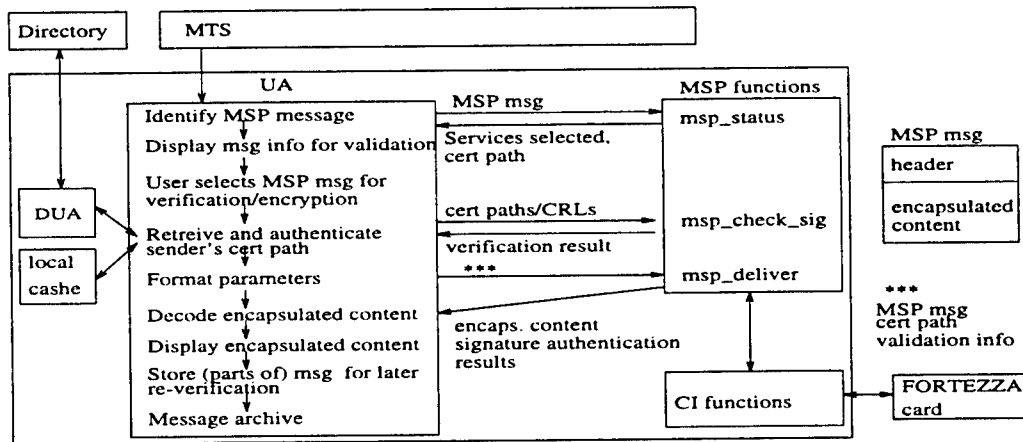


Figure 6: MISSI Receive Mail

From [5], [4] and [2], we specify receiver algorithm as follows:

Receive new message

UA informs the user about the received message

user selects message for processing

UA checks if the complete msg is signed;

if yes, validate sequence signature

UA calls msp_status to examine the message

msp_status returns msg parameters, including:

security services applied

certification paths: sequence signer's

message signer's

message encryptor's

message classification
message content, if the message was not encrypted

UA verifies each certification path returned by msp_status

UA displays to the user:

content description
services applied
authenticated identity of msg signer
authenticated identity of msg encryptor
security label (if msg was encrypted)

UA calls:

msp_deliver to process incoming msg; or
msp_val_receipt, to process incoming receipt

if msg was encrypted, msp_deliver will:

select token
perform prbac
if prbac passes,
extract lrbac information from the token
perform lrbac
if lrbac passes,
decrypt msg
create signed-only msg version for forwarding/revalidation

if msg was signed, msp_deliver will:

perform signature validation
perform receipt request processing
receipt may be returned automatically or after confirmation
from user

If msg was not encrypted,

UA saves it for forwarding/revalidation

If the encapsulated content contains forwarded messages,

UA must extract them and manage them for future forwarding and revalidation.

UA disposes message, or

retains it for storage and revalidation:

store the original, or
store content only, or
store content and msg signature, for forwarding and non-repudiation

If a user requests that the message encapsulated by a stored message be
displayed but not validated,

UA calls msp_status to provide the encapsulated content

If the user requests that the message be validated,
UA calls `msp_deliver` or `msp_val_receipt`

7.1 `msp_status`

`msp_status` is used to:

- decode, but not decrypt incoming message
- determine security services and message attributes: content description, content type and organizational flag
- extract certification paths

The inputs to `msp_status` include:

`msp_path`: pointer to the ASN.1 encoded message

`content_path`: pointer to the file in which to store encapsulated content, if the message was not encrypted

The outputs of `msp_status` include:

`content_path`

`serv_sel`: services applied to the message

`msg_atts`: message attributes: content description, content type and organizational flag

`org_paths`: certification paths of sequence signer, message signer, and key exchange token signer (i.e. message encryptor)

7.2 `msp_deliver`

`msp_deliver` is used to:

- provide encapsulated content (i.e. decrypted message if the message was encrypted), services applied, organizational flag, and other information for incoming messages, after the messages have gone through `msp_status`. `msp_deliver` will also verify sequence signature and message signature, and partially re-authenticate certification paths.
- create signed-only version of the incoming message (to be used for forwarding and revalidation), if the message was encrypted.
- create signed receipt for the incoming message, if requested

The inputs to `msp_deliver` include:

`msp_path`: pointer to the pathname for the incoming ASN.1 encoded message

cert_index: certificate number for the user-selected personality

content_path: pointer to pathname in which msp_deliver will store the unencrypted encapsulated content

org_paths: pointer to originator certification path structure:

seq_cert_path: pointer to the certification path of sequence signer

msg_cert_path: pointer to the message signer

key_cert_path: pointer to the message encryptor certification path

sign_cert_path: pointer to the file which will contain ASN.1 encoded, MSP formatted signed receipt

fwd_path: pointer to the file which will contain the ASN.1 encoded, MSP formatted message containing the signature block and unencrypted encapsulated content from the incoming message, if it was signed and encrypted.

fwd_seq: flag indicating if the signed-only MSP message saved for forwarding and revalidation (in fwd_path) should get sequence signature applied to it.

rec_seq: flag indicating if sequence signature must be applied to the signed receipt generated for the incoming message.

The outputs from msp_deliver include:

serv_sel: which security services have been applied to the message

content_path: unencrypted, encapsulated content

fwd_path: signed-only version of the message, stored for forwarding/revalidation

msg_atts: message attributes: security label, content description, organizational flag, content type, security categories, and policy ID.

8 Forwarding

[4] p.14 specifies:

“20. Allow the user to forward the following:

- a. Unprotected messages
- b. MSP signed messages with message signature
- c. The verified content of MSP protected messages.”

Requirement 20.c is saying that a user is allowed to forward decrypted contents of formerly encrypted message, without any signatures to prove where the original signed and encrypted message came from (encrypted messages are always signed). This can be a security hole.

21, 22. Allow the user to receive, review, and reply to unprotected and MSP protected messages.

23. Allow the user to receive and review to unprotected and MSP protected messages that contain forwarded MSP signed messages with message signature.”

Requirements 20 and 23 seem to contradict each other - a user is allowed to forward three kinds of messages, yet is allowed to receive and review only one kind of the three.

What happens if user A forwards to user B an unprotected message or decrypted contents of formerly encrypted message. According to requirement 23, user B is not able to receive and review the message.

“24. Be capable of inserting file attachments into messages that will be unprotected and MSP protected(,) and extracting file attachments from unprotected and MSP protected messages.”

There are no other restrictions on file inserts. Since all MISSI messages can be stored as the decrypted content only, as specified in clause 20., this means that we can send contents of any MISSI message as a file inclusion in an unprotected message. This is a security hole.

“4. Allow the user to change Card individual and organizational personalities during a session.

a. When displayed to the user, the application shall interpret the four byte certificate Usage/Equipment specifier as follows: “INKS” as “Individual,” “INKX” as “Individual read only,” “ONKS” as “Organizational,” and “ORKX” as “Organizational Read Only.”

b.

c. ... ”

“5. Allow user to logout and re-login to the Card at any time without having to restart the application.”

Each user has only one PIN, which is used to allow access to FORTEZZA Cards.

Each user has multiple personalities, and each personality has its own certificate and distinguished name (DN).

However, [5] p.11 forwarding specifications conflict with [4] clause 20: “An MSP message may be retained in at least three forms:

- original form (exactly as it arrived prior to any processing)
- content only (without any MSP heading information)
- content with signature (content retaining the message signature information to provide non-repudiation and to support forwarding).”

[5] specifies forwarding rules as:

- any number of forwarded MSP messages may be conveyed within a new message
- forwarded MSP messages may be nested within one another
- messages that can be forwarded are:
 - signed receipts

- the original unencrypted content and the original MSP signature
- encrypted MSP messages (if the recipient possesses the cryptographic material required to decrypt the forwarded MSP message)
-

[17] p. A-28 specifies the forwarding rules as:

- original encrypted content type and envelope information
- the original contents and delivery information (encrypted messages will be decrypted prior to forwarding). *Are the signatures preserved? Or we send the original content without any signature?*

MSP allows for forwarding MSP messages within a MSP message. ACP 123 allows for forwarding MM or IPM within a MM message

9 Secure Network Server (SNS)

Only Secret enclaves are required to have SNS.

[4] p.7 says: "SNS for Secret High Enclave will:

- do positive identification and authentication for messages in and out of the enclave in accordance with local security policy
- check for proper invocation of security services for messages leaving the enclave in accordance to the local policy
- regrade messages according to the local policy:
 - downgrade messages destined for lower level enclaves
 - upgrade messages entering from lower level enclaves

How does SNS know what enclave level a message is coming from? Message headers carry information about individual messages, not enclaves.

The above requirement is not clear - does "destined to lower level enclave" means that a message is leaving the present enclave and going to a lower level enclave; and "entering from lower level enclave" means that a message originated at a lower level enclave is entering from WAN.

- reject or admit unprotected unclassified messages according to the local policy."

9.1 Omission in CONOP

An important omission is present in [14] p.23, because message headers carry information about individual messages, not enclaves. Therefore, SNS has no way of determining the security level of the enclave which originated the incoming message, unless such information is stored in SNS. SNS can be easily configured to “know” the security level of the enclave it belongs to. Security levels of all other enclaves would have to be stored in some table within SNS. Such tables would be analogous to routing tables of Internet gateways. However, enclave security level can change in time, and there has to be a way to update the tables of all SNSs. Both the presence of tables and updating process need to be included in SNS specification, or local Directory specification. We think that SNS or local DUA would be the most logical place to include enclave security level information.

This omission probably happened because CONOP includes only example of messaging between Secret High enclaves.

Enclaves may consist of workstations of various security levels: “A Secret System High enclave contains at least one FORTEZZA protected user workstation. It may also contain workstations that are not FORTEZZA protected, .. ” [14] p.20. MISSI local policy depends on enclave security level, as we discussed in detail in section 10.

[14] Appendix D, states: “When MSP encrypted/signed messages are processed at SNS, the SNS will remove its token from the message prior to delivering the message. If a sequence signature has been applied to the message by the originator, then this outer signature is invalidated by the removal of SNS token.

Thus, the operational scenarios **assume** that if the MSP optional sequence signature is selected by the originator, then the use of a sequence signature on that specific message must be preserved throughout the transmission from writer to reader.

Thus, SNS must generate a new sequence signature for the message in order to preserve the use of sequence signature.” *Is this SNS on sender’s or receiver’s side?*

“Other assumptions are: X.400 services will not be used to transfer a X.400 P772 message from the workstation to the SNS. SNS has MTA functions that can transfer messages (for example, P1 envelope) out of the enclave. These assumptions are based on understanding that current MTA products do not completely support the MTA functionality (e.g. P3 submit envelopes) specified in X.400. With the MTA functionality in the SNS, multiple MTAs will not be required in the enclave.”

As an outbound filter, i.e. when checking messages which are leaving the (Secret) enclave, SNS will pass messages according to the following criteria [14] p.73: all S messages must be encrypted; all messages must be signed.

SNS will also check if the message originator is allowed to send messages with security level lower than the enclave security level.

As an inbound filter, i.e. when checking messages which are entering the enclave, SNS will accept messages according to one of the following three policies [14] p.79: messages must be signed;

or messages must be encrypted; or any messages are accepted.

10 Relationship Between MISSI Certificates, Workstations, Message Security Levels and User Authority Properties

In order to formalize MISSI access policy, we put together a more formal definition of terms and rules scattered throughout MISSI documentation.

The term FFC means "FORTEZZA for Classified."

Each user has one PIN number that allows access to his FORTEZZA card.

Each user, or PIN, has several personalities.

Each personality has its own DN and certificate stored on the card.

The relationship between user, Card, personalities and certificates is shown in Fig. 7.

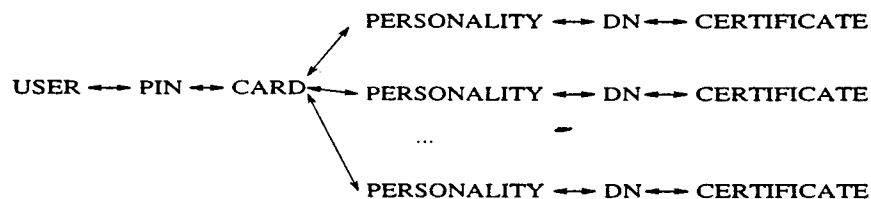


Figure 7: Relationship Between User and His Personalities

When a user logs into his Card using a chosen personality, the Card will perform all operations from that point on using the certificate associated with the chosen personality.

Once a user is logged into Card, he can change personalities any time during a session.

User can logout and re-login to the Card without logging out of the host [4], p.2. *This requirement could imply that a non-FORTEZZA user can login into a workstation and use the workstation as a plain-vanilla workstation, if there is no additional mechanism to restrict workstation use to FORTEZZA users only. This could be a serious security hole in case of Secret workstations.*

There are separate certificates for SBU and S personalities. SBU and S certificates are distinguished by cryptographic universals used to create the certificate. X.509 certificates for Secret are only capable of encrypting messages for users also possessing an X.509 certificate for Secret [14] p.28. *Therefore, SBU certificate cannot be used to decrypt information encrypted by S certificate, and vice versa.*

The latest CONOP [14] is not very precise in terms of certificate requirements. It says: "If a S user requires the capability to originate traffic destined for an SBU user outside of the S enclave or

to receive messages from an SBU only user, then the user FORTEZZA card must host two X.509 certificates: one for encrypting S traffic and one for encrypting SBU traffic." "SBU/Secret Dual Mode Card contains one or more X.509 certificates for Secret as well as one or more X.509 certificates for SBU. This card is capable of originating and receiving both Secret and SBU messages depending on the classification of the data within the message and the capabilities of message recipient." *We assume that this means that the S user must have two personalities: S and SBU, and switch personalities when he or she needs to send or receive encrypted messages. Perhaps this is why a card allows personality switch within a login session. Another interpretation of the above requirement would be that the dual mode card automatically switches personalities when needed; for example, a S user receives a message from SBU user, so the receiver's card automatically uses the SBU certificate associated with the S user - but this interpretation does not make sense from the security standpoint.*

"Since the universal within the signature component of the FFC X.509 certificate is the same as in the certificate for Secret and certificates for SBU, users with only one certificate (e.g. DSA with only a certificate for SBU) are capable of verifying signatures of directory accesses using either the certificate for Secret or the certificate for SBU." [14] p.28. This requirement does not make sense if read literally, because it says that a user possessing only SBU certificate can use either SBU or S certificate - but it is stated that the user has only SBU certificate and therefore cannot use S certificate which he does not possess. However, if we try to interpret this requirement, we can assume that it means: *Users with only one certificate are capable of verifying signatures signed using either S or SBU certificate. Therefore, signatures signed using S or SBU certificates can be verified using either S or SBU certificates. In other words, S personality can read SBU signature, and vice versa.*

"Certificate for S has S, SBU and U privileges, and certificate for SBU has SBU and U privileges" [14] p.61. We think that this statement means that personalities with S certificates are allowed to send messages classified as S, SBU or U; and personalities with SBU certificates are allowed to send messages classified as SBU or U.

There are separate, dual certification hierarchies for S and SBU certificates. Therefore, there are PAA, PCAs and CAs for SBU certificates, and PAA, PCAs and CAs for S certificates [13] p.16.

There are several classes of FORTEZZA cards: SBU only, which contain only SB certificates; S only, which contain only S certificates; dual, which contain SBU and S certificates; SBU traveling, and compartmented only card. Both dual and Secret only cards are called FFC cards [14] p.47.

Each workstation has its own level of security level (we conclude this from SDN.701, p.8). We assume this security level must be based on which FORTEZZA card reader is connected to the workstation: U if there is no card reader, SBU if there is FORTEZZA for Sensitive But Unclassified Card reader (FFS), and S if there is FORTEZZA for Classified (FFC) Card reader. We see from [14] p.46 that CA workstation can have up to three PCMCIA slots: one slot for the Dual Use Cards (S/SBU), one slot for S Cards only, and one slot for SBU Cards only. We assume that other workstations have the same configuration.

Enclave security level is equal to the highest security level of the workstations which form the enclave. An enclave can contain workstations of lower security level. For example, an enclave is proclaimed Secret if it contains at least one FFC user workstation.

A Secret enclave can contain FFS workstations as well as workstations with no FORTEZZA Card readers. We conclude this from the sentence: "A Secret System High enclave contains at least one FORTEZZA protected user workstation. It may also contain workstations which are not FORTEZZA protected, ..." [14] p. 20.

Currently, we do not see anywhere in MISSI documents that workstations are aware what is the security level of the enclave they are in.

Each personality is free to log into any workstation which is connected to a FORTEZZA Card reader of the level higher or equivalent to the personality's security level. For example, S personality cannot login into an SBU capable card reader. (Because S and SBU certificates are different).

We assume that a card reader of security level X can read cards of security level X and below.

It follows that a Secret personality must use a FFC capable workstation, therefore it must operate from a Secret enclave.

[14] specifies an access policy between users in Secret enclaves and users at other enclaves. An example is: "Between Secret enclaves: signed and encrypted unclassified through Secret (U-S) messages originated by users at FFC workstations in Secret enclaves for recipients at FFC workstations in Secret enclaves." This requirement is ambiguous. The questions we need to ask are:

- the word "user" should be replaced with the word "personality."
- *Is it possible to log into a workstation if a user is not a FORTEZZA user, or has authorization level less than the Card reader level for that workstation?* That is, can a SBU personality use the workstation which has FFC Card reader connected to it? If it is possible, then the above requirement must be modified. The above requirement is valid only if Secret personalities exclusively can use a FFC workstation.

From the MISSI documentation, it is not clear if the above holds. However, from talking with Chet Hossmer from ORA, who worked on MISSI implementation, the above requirement holds; therefore, the CONOP is incorrect.

Therefore, we conclude that MISSI access policy has to be developed as a matrix which takes into account security levels of: enclave, workstation, and personality currently used, as shown in Fig. 9. The matrix needs to be filled with the allowed message characteristics: message classification and security service applied to the message. We filled the matrix with a security policy we proposed and tested in our model. We summarized various security levels in section 11.

| SECURITY SERVICE APPLIED | MESSAGE SECURITY LEVEL | | | |
|---------------------------|------------------------|------|------|----------------------|
| | U | SBU | S | |
| UNPROTECTED (UP) | 11 ✓ | 12 ? | 13 X | Vanila workstation |
| SIGNED ONLY (SO) | 21 ✓ | 22 ✓ | 23 X | FOTREZZA workstation |
| SIGNED AND ENCRYPTED (SE) | 31 ✓ | 32 ✓ | 33 ✓ | |

← vanilla user →

← SBU user →

← S user →

✓ : POSSIBLE COMBINATION

X : IMPOSSIBLE (I.E. FORBIDDEN) COMBINATION

? : NOT CLEAR IF IT IS A POSSIBLE COMBINATION

Figure 8: MISSI Message Security Classifications

Therefore, we formalize the acces policy as follows: Any personality of security level X is allowed to send email to any other personality of security level Y, if:

- the messages sent are classified at the security level Z, where $Z \leq \min(X, Y)$;
and
- the messages sent have security service W applied to them, where W is determined as:
 - we do not know if SBU enclaves are allowed to send unprotected messages to the WAN - this is not discussed in [14].
 - all messages classified as Secret must be signed and encrypted; i.e. $W = \text{signed} + \text{encrypted}$ if $\text{msg_level} = S$.
 - Secret enclaves cannot send unprotected messages.
 - Secret enclaves cannot receive unprotected messages. That is, $W \geq \text{signed}$ if $\text{wkst_level} = S$. In other words, Secret enclaves must send and receive signed or signed-and-encrypted messages. This requirement is enforced by SNS filtering [14]. p.11.

This definition can cause problems. For example, a SBU personality can be working from a Secret enclave, from a SBU capable workstation. Currently, we do not see anywhere in MISSI documents that workstations are aware what is the security level of the enclave they are in. This personality could send an unprotected unclassified message to a user on a vanilla enclave or workstation.

We conclude that all workstations in the Secret enclave must have the same UA implementation, which does not allow unsigned messages. However, in a Secret enclave we can have vanilla workstations and non-MISSI users, which can send unprotected messages. The only way to prevent non-protected messages leaving a Secret enclave is to have the Guard know enclave level and prevent unauthorized messages from leaving the enclave.

[11], p.14, specifies that "UAs are grouped into classes based on the type of content of messages they can handle. The MTS provides a UA with the ability to identify its class when sending messages to other UAs."

This requirement means that users at vanilla workstations in Secret enclaves cannot send anything outside the enclave (provided that SNS filtering works).

- vanilla workstations can receive only unprotected messages, because they do not have any FORTEZZA capability.*

The last two requirements clash, because if a Secret enclave sends a signed message to an unprotected enclave, the unprotected enclave cannot verify the signature, because an unprotected enclave does not have a Card reader yet it needs `msh_check_sign` function from the Card to verify the signature.

An old version of CONOP proposes that secret enclaves can send unprotected messages to SNS, which will archive the message content for future audit. The new CONOP [14] addresses this issue implicitly. It assumes that a Secret enclave sends a signed message to an unprotected workstation, and that this message consists of two identical parts, one with the signature and another with no protection. SNS will verify the signature and assure that the two message parts are identical [14] p.7. 72. We conclude that, in this case, the SNS on the receiver's side must know the capabilities of the recipient, and remove the signature if the recipient has no Card reader.

The above discussion brings us to formalize one additional requirement:

- *secret enclaves can communicate with vanilla workstations only if the vanilla workstations are on an enclave with SNS.*

Currently, only Secret enclaves are obliged to have SNS.

If a Secret workstation sends a signed message to a vanilla workstation which is not in an enclave with SNS, the receiver can be used to collect messages from the secret enclave, which presents a security hole.

Secret Messages between a Secret personality and any other personality must be signed and encrypted. SBU Messages between a Secret personality and any other personality are either signed and encrypted, or signed only.

11 Security Levels

In this section we will clarify various security levels: enclave, card, workstation, user, certificate, and message.

11.1 Enclave Security Level

If an enclave E consists of workstations W1, W2, ... Wn, and the workstations have security levels S1, S2, ... Sn respectively, then the enclave E has security level SE, where $SE = \max(S1, S2, \dots, Sn)$.

11.2 Workstation Security Level

Workstation security level is determined by the capability of the Card reader attached to the workstation. [14] mentions different “slots” on card readers. Each “slot” accepts Cards of certain security level. a Card reader can have the following slots: a slot capable of reading only SBU-only Cards, a slot capable of reading only S-only Cards, and a slot capable of reading Dual-use SBU/S cards (these Card readers are called Dual SBU/S Card readers). From the wording used in [14] p.43 and p.46, it seems that a card reader with a Dual slot can have another slot used only for S-only or SBU-only cards.

11.3 Card Security Level

Card security level is determined by the security level of certificates and cryptographic and signature verification material stored on the Card.

There are three kinds of cards: S-only, SBU-only, and Dual S/SBU. S-only Cards contain only S certificates and materials, SBU-only Cards contain SBU certificates and materials, and Dual S/SBU Cards contain both S and SBU certificates and materials.

Each card belongs to a particular user. A user can own several certificates, which are stored on the user's Card. The user unlocks the Card by supplying his PIN.

11.4 Certificate Security Level

There are two kinds of certificates: SBU and S. Certificate for S has S, SBU and S privileges (i.e. can perform security services on messages classified as S, SBU or U), and certificate for SBU has SBU and U privileges.

Certificates are also classified as organizational or personal.

Based on certificate security and organizational classification, the user can or cannot perform certain operations, based on local security policy and other MISSI requirements.

SBU and S certificates have different cryptographic material, so the sender and the receiver must use exclusively either S or SBU certificates in order to communicate encrypted messages.

SBU and S certificates have the same signature verification material, so SBU and S certificates are interchangeable for signature verification.

11.5 User Security Level

User security level is determined during a MISSI session by the type of Card this user possesses and the certificate the user is currently using. Each user can have several identities, called personalities, where each personality owns a certificate and has a certain security level and organizational status assigned to it. For example, Joe Schmoe can have a Card with two certificates: S organizational certificate as Joe Schmoe the Division Director, and SBU-only personal certificate as Joe Schmoe an employee. Certificates are stored on the users personal Card (and unlocked via the user's personal PIN). Each user can switch personalities during a MISSI session.

11.6 Message Security Level

Message security level is determined by the local security policy. Either the user is allowed to assign message security level, or UA assigns it based on the local security policy. For example, the local UA can assign the message security level based on the workstation or enclave security level [14] p.59.

ACP 123 specifies that each military message must have (a clearly displayed) security classification. [16] p.4-9 determines the message security label as: "The security label assigned to the entire message will be that of the highest classification of any part of the message or the appropriate label for the aggregate of the information contained in the entire message including all body parts." Each body part can be a forwarded message.

11.6.1 Body Parts

If a message M consists of message text T and several body parts $B_1, B_2, \dots B_n$, and if the user assigns security level ST to the message text, and body parts have security levels $S_1, S_2, \dots S_n$ respectively, then the message M has security level SM , where $SM = \max(ST, S_1, S_2, \dots S_n)$.

An MM body part can contain forwarded-MM and forwarded-IP messages.

12 SPIN Specification of MISSI

In order to utilize model checkers, it is necessary to have a manageable number of states, which leads to attempts to save on the number of states. There are several ways to reduce number of states: to bundle several consecutive execution steps in one unit using `atomic` statements; to reduce channel sizes, counters, assert and other statements, and number of variables in general; and to make an abstraction of the system and focus on the areas of interest only.

We employed all these approaches in our modeling. Our model consists of modules, where we abstract the modules which are less important by their input-output function, and focus on the modules of interest.

Currently, we have three different models of MISSI: [6] specifies a detailed model of MISSI sender with an abstraction of local cache and certificate verification; [7] also specifies the sender, but focuses on the local cache and certificate verification, and abstracts away the processes necessary to prepare a message for sending. In this report, we abstract away the sending process, and focus on the receiver.

13 Specification of MISSI Receiver and Forwarding

We model a network consisting of two workstations connected to the WAN via SNS, as shown in Figure Fig. 10. We specify receiver algorithm and incorporate descriptions of SNS and MSP functions as outlined in sections 9 and 7.

Workstation A is the sender, and B is the receiver. We simplified the sender, in order to eliminate unnecessary states. We have a detailed model of the sender and message verification process in [6] and [7]. In this report, sender model details only the structure needed to assign message classification and security services and chose recipients. We assume that the message passes verification tests and preparation for sending.

On the receiver side, we assume that the message passes verification tests, and we outline the steps needed to receive the message.

We have two receiver models: in the first model, which we called Model 1, we assume that the sender can forward any message it has stored previously. In Model 2, we assume that the receiver can forward any message that the sender has sent. We had to split the forwarding capabilities into two models in order to reduce state space and be able to run the model. We did merge capabilities of

Model 1 and Model 2 together, but this composite model would always run out of memory, because the network channels are quite large with the addition of a forwarded message (which adds three fields to the channels). Therefore, we decided to split the capabilities.

We took several additional measures to reduce number of states. We “hard-wired” the two enclaves together, in order to eliminate routing states. We did not model calling any of the MSP functions except `msp_status` and `msp_deliver`, which are needed for receiving. We assume that all other functions are called and are executed with no errors. We include them in the model as comments. In Model 2, we do not include any MSP functions, not even `msp_status` and `msp_deliver`. We do model them in Model 1. Model 1 was at the limit of available memory space. SPIN would run out of memory if we tried to compile the model as `pan -w25 -m400000`, i.e. to limit the number of states to 2^{25} . We did compile the model with `-w28`, which is the upper limit on a workstation with 320Mb of RAM.

14 Variables

We use the following variables, with their values in parenthesis:

`wkst_capability`: workstation security level (S, SBU, S, or Dual)

`enclave_level`: enclave security level (S, SBU, or U)

`pers`: personality used; also what type of certificate was used (organizational SBU and S personalities/certificates are called `ORG_SBU_PERS` and `ORG_S_PERS`; individual SBU and S personalities/certificates are called `SBU_PERS` and `S_PERS`; and non-MISSI user with no certificate is called `U_PERS`)

`to`: personality of the receiver

`from`: personality of the sender

The following variables refer to the overall message sent/received:

`msg_classification`: message security level (S, SBU, or S)

`serv_sel`: security services applied to the message (NONE, SIGN, ENCRYPT)

`org_msg`: organizational message flag

The following variables refer to the message body part that contains forwarded message:

`forwarded_msg_clas`: message security level (S, SBU, or S)

`forwarded_serv_sel`: security services applied to the message (NONE, SIGN, ENCRYPT)

`forwarded_org_msg`: organizational message flag

15 Modeling Messages

Since SPIN cannot pass arrays as channel parameters, we modeled message encapsulation and “peeling” of message headers indirectly. We passed header fields of the message as channel arguments, and assumed that those arguments which are “hidden” are not accessible because the message was

not stripped to that part yet. Those arguments which are “visible” are those which are in the header that is currently used.

For example, the user sends message MSG, which contains body parts BODY1 and BODY2. We add header HDR to the message. The header contains fields MSG_CLAS, ORG_MSG, and SERV_SEL. MSG and HDR together make an MSP message. This message is enveloped by X.400 header X400_HDR, which contains FROM and TO fields, and the entire MSP message becomes the body of the X.400 message. We would like to model sending of this X.400 message as:

```
to_net!X400_HDR, X400_body
```

where X400_body would be an array consisting of arrays HDR and MSG. However, this is not possible in SPIN. Sending of this X.400 message will be modeled as:

```
to_net!FROM, TO, d1, d2, d3, d4, d5, d6, d7, d8, d9, d10, d11
```

FROM and TO fields are “visible,” i.e. they are used for the current routing and other processing. d1, ..., d7 are “invisible” dummy variables which contain the parts of the message which are currently enveloped and inaccessible. We need to “peel off” the headers in order to reach d1, ..., d7. d1 represents field MSG_CLAS, d2 represents field ORG_MSG, d3 represents SERV_SEL, d4, d5, d6 and d7 represent MSG_CLAS, ORG_MSG, SERV_SEL and BODY parts of BODY1 (which is a complete message in itself); and d8, d9, d10 and d11 represent BODY2.

16 SNS Model

The outbound messages are examined as:

```
if
:: (d5 <= from && d5 <= to) -> skip;
  if
  :: (d5 == S && d6==SIGN+ENCRYPT) -> skip;
    if
    :: (enclave_level == S && d6 >= SIGN) -> skip;
      ::!(enclave_level == S && d6 >= SIGN) -> wrong ++;
        printf("wrong=%d\n", wrong); goto end_net;
    fi;
  ::!(d5 == S && d6==SIGN+ENCRYPT) -> wrong ++;
    printf("wrong=%d\n", wrong); goto end_net;
  fi;
::!(d5 <= from && d5 <= to) -> wrong ++;
  printf("wrong=%d\n", wrong); goto end_net;
fi;
to_wan[s]!from, to, fdb1,fdb, fd5, fd6, db, d5 , d6;
```

Inbound messages are examined as:

```

:: to_wan[r]?from, to, fdb1,fdb, fd5, fd6, db, d5 , d6;
atomic {
    if
    :: (enclave_level == S && d6 >= SIGN) -> skip;
    ::!(enclave_level == S && d6 >= SIGN) -> wrong ++;
        printf("wrong=%d\n", wrong); goto end_net;
    fi;
    from_net[wkst]!from, to, fdb1,fdb, fd5, fd6, db, d5 , d6;
}

```

17 PRBAC Model

In case the message was encrypted, we perform PRBAC. We assume that LRBAC will pass.

```

prbac:
if
::(msg_classification <= pers && msg_classification <= from ) -> skip;
    if
    :: ( msg_classification <= wkst_capability) -> skip;
    ::!( msg_classification <= wkst_capability) ->
        msg_classification = wkst_capability;    /*upgrade*/
    fi;
    ::!(msg_classification <= pers && msg_classification <= from) ->
        wrong ++; goto exit_;
fi;

```

18 Personality Switching

We model that receiver can switch personality during a session, in order to receive a newly arrived message:

```

if
:: (to == pers) -> skip;
::!(to == pers) ->
    if
    :: (wkst_capability == U && to == U_PERS) ->
        pers = to; /*switch personalities*/
    :: (wkst_capability == SBU && (to == SBU_PERS || to == ORG_SBU_PERS || to == U_PERS)) ->
        pers = to;
    :: (wkst_capability == S && (to == S_PERS || to == ORG_S_PERS || to == U_PERS)) ->
        pers = to;
    :: (wkst_capability == Dual) ->
        pers = to;

```

```

::!( (wkst_capability == U && to == U_PERS)
|| (wkst_capability == SBU && (to == SBU_PERS || to == ORG_SBU_PERS || to == U_PERS))
|| (wkst_capability == S && (to == S_PERS || to == ORG_S_PERS || to == U_PERS) )
|| (wkst_capability == Dual)) -> wrong++;
assert(received_mail == sent_mail || wrong == 1);
goto exit_;
fi;
fi;

```

19 Model 1

In Model 1, we assume that the sender can pick up any stored message and forward it. This is equivalent to forwarding a message from Pine mailbox.

```

if
:: skip;          /*do not forward anything*/
:: forward = 1; /*pick and send any existing message*/
if
:: forwarded_msg_clas = S;
:: forwarded_msg_clas = SBU;
:: forwarded_msg_clas = U;
fi;
if /*this should not be possible for U_PERS or wkst_level =U */
:: forwarded_org_msg = 1;
:: forwarded_org_msg = 0;
fi;
if
::!(pers == U_PERS) ->
if /*this should not be possible for U_PERS or wkst_level =U */
:: forwarded_serv_sel = ENCRYPT + SIGN;
:: forwarded_serv_sel = SIGN;
:: forwarded_serv_sel = 0;
fi;
::(pers == U_PERS) -> skip;
fi;
fi;

```

20 Model 2

In Model 2, we assume that receiver is forwarding a freshly received message. The receiver is operating from personality `pers`.

```

if
:: (pers == ORG_SBU_PERS || pers == ORG_S_PERS ) ->

```

```

        -> forwarded_org_msg = 1;
    /*organizational users can send only organizational msgs*/
    ::(! (pers == ORG_SBU_PERS || pers == ORG_S_PERS) && org_msg == 0) ->
        forwarded_org_msg = 0;
    :: CASE1 or CASE2
    fi;

/* CASE 1: LIMITED FORWARDING OF ORGANIZATIONAL MESSAGES*/
/*assume that non-org user CANNOT forward org msgs*/
::(! (pers == ORG_SBU_PERS || pers == ORG_S_PERS) && org_msg != 0) ->
    goto finish_receiving;

/* CASE 2: UNLIMITED FORWARDING OF ORGANIZATIONAL MESSAGES*/
::(! (pers == ORG_SBU_PERS || pers == ORG_S_PERS) && org_msg != 0) ->
    if
    ::(forwarded_serv_sel == 0) -> forwarded_org_msg = 0;
        /*since the forwarded msg is not signed, it will not be
        recorded as organizational*/
    ::!(forwarded_serv_sel == 0) -> forwarded_org_msg = 1;
    fi;

```

21 Results

We ran our SPIN model on an Indigo-2 with 320Mb of RAM. The model took several seconds to execute.

If we forward a message, we assert that the resulting message classification must be greater than the individual parts:

```
assert(forwarded_msg_clas <= msg_classification);
```

If we forward a message, we assert that the resulting message must be organizational if the original was organizational:

```
assert(forwarded_org_msg <= org_msg);
```

We also keep counters `sent_mail` and `received_mail`, which are incremented every time mail is sent or received. In case SNS or the receiver rejects a message, we increment counter `wrong`. We assert that either we receive correctly, or the message is lost:

```
assert((received_mail == sent_mail && wrong == 0)
|| (received_mail == sent_mail && wrong == 1));
```

```
indigo > spin -a main.spin16
```

```
indigo > gcc -DMEMCNT=28 -DBITSTATE -DSAFETY -o pan pan.c
indigo > pan -w24 -m300000
```

We ran our model for Model 1, where only sender can forward, and for Model 2, where only receiver can forward.

21.1 Model 1

All assert statements are true.

```
indigo > pan -w28 -m400000
(Spin Version 2.9.4 -- 4 November 1996)
+ Partial Order Reduction
```

Bit statespace search for:

```
never-claim          - (none specified)
assertion violations  +
cycle checks          - (disabled by -DSAFETY)
invalid endstates     +
```

State-vector 536 byte, depth reached 102, errors: 0

8.28066e+06 states, stored

1.30117e+06 states, matched

9.58183e+06 transitions (= stored+matched)

7.67589e+06 atomic steps

hash factor: 32.4172 (expected coverage: >= 98% on avg.)

(max size 2²⁸ states)

4.44963e+09 equivalent memory usage in bytes (stored*vector + stack)

7.83458e+07 actual memory usage

unreached in proctype Guard

line 67, state 47, "-end-"

.....

21.2 Model 2: Unlimited Forwarding of Organizational Messages

If we allow unlimited forwarding of organizational messages, `assert(forwarded_org_msg <= org_msg)` does not hold.

For example, assume that an organizational user, let us call him A, sends a signed and encrypted message to a non-organizational user B. B saves the original encrypted message and tries to forward it. Let us represent the original message as OM. B constructs a new message NM, which will be used to forward OM. Since B is not an organizational user, he cannot mail an organizational message; therefore, we have an organizational message OM embedded in a non-organizational message NM. This embedded message is still encrypted, which provides some protection - unless B sends his key to the recipient of the forwarded message.

The same scenario could have happened with signed-only original organizational message OM.

pan: assertion violated (forwarded_org_msg<=org_msg) (at depth 104)

pan: wrote main.spin16.trail

(Spin Version 2.9.4 -- 4 November 1996)

Warning: Search not completed

+ Partial Order Reduction

Bit statespace search for:

| | |
|----------------------|--------------------------|
| never-claim | - (none specified) |
| assertion violations | + |
| cycle checks | - (disabled by -DSAFETY) |
| invalid endstates | + |

State-vector 448 byte, depth reached 107, errors: 1

331578 states, stored

62288 states, matched

393866 transitions (= stored+matched)

648730 atomic steps

hash factor: 50.5979 (expected coverage: >= 98% on avg.)

(max size 2²⁴ states)

1.56947e+08 equivalent memory usage in bytes (stored*vector + stack)

1.26272e+07 actual memory usage

21.3 Model 2 Correction: Limited Forwarding of Organizational Messages

If we allow limited forwarding of organizational messages, `assert(forwarded_org_msg <= org_msg)` holds.

(Spin Version 2.9.4 -- 4 November 1996)

+ Partial Order Reduction

Bit statespace search for:

| | |
|----------------------|--------------------------|
| never-claim | - (none specified) |
| assertion violations | + |
| cycle checks | - (disabled by -DSAFETY) |

```

invalid endstates      +

State-vector 448 byte, depth reached 109, errors: 0
  433240 states, stored
    85051 states, matched
    518291 transitions (= stored+matched)
    855639 atomic steps
hash factor: 38.7249 (expected coverage: >= 98% on avg.)
(max size 2^24 states)

2.02492e+08    equivalent memory usage in bytes (stored*vector + stack)
1.26272e+07    actual memory usage

unreached in proctype Guard
.....

```

It is necessary to prohibit non-organizational users from forwarding organizational messages. Therefore, we propose the following requirement: *Every message sent must have the following property: If a message M consists of message text T and several body parts $B_1, B_2, \dots B_n$, the message M is assigned organizational status iff any of the body parts has organizational status and it is possible to assign organizational status to M . If it is not possible to assign organizational status to M , the message cannot be sent.*

22 Conclusion

In our experience, English specifications are ambiguous and incomplete. Even specifications that sound correct can lead to incorrect results. At least, English specifications should contain some kind of pseudo-code or flow diagram, in order to see more clearly all "if-then-else" branchings and other possibilities.

Protocol specification usually relies on other standards and protocols, as MISSI relies on X.500, X.400, ACP 123 and SDN.701, and the interface needs to be clearly specified.

Formal specifications and proofs, augmented with textual descriptions, provide a very clear and detailed documentation for the system, aid in understanding of the system, and expose weak areas of system design. Formal specification can be used as a blueprint for implementation, and formal proofs can be used as a blueprint for implementation testing.

Formal verification assures of desired system properties. Desired system properties consist of original specification, enhanced by additional requirements that formal specification uncovers.

Formal specification must contain some implementation bias. Specifiers must be careful to minimize it.

Model checkers always eventually run out of memory. In order to run a model, we must divide it in several sub-models, and validate each sub-model. The sub-models can represent the same system,

where each sub-model emphasizes a particular feature of the system and simplifies the other features (e.g. one sub-model focuses on login and logout, another on sending mail, another on receiving mail). The sub-models can be parts of a bigger model (e.g. each submodel is a layer in a layered protocol.) In this work, we used the first approach.

SPIN is useful as a tool, especially because it resembles C.

Future work includes further formalization of the required properties into a form that can be machine checked. We plan to use either first or higher order logic, which can be verified using Penelope or HOL.

References

- [1] *Interface Control Document for the FORTEZZA Message Security Protocol Software*. http://www.armadillo.huntsville.al.us/Fortezza_docs/obsolete.html, December 20 1994.
- [2] *FORTEZZA Application Developers Guide*. Publicly available documents at <http://www.armadillo.huntsville.al.us>, 14 July 1995.
- [3] *FORTEZZA Simple Mail Transfer Protocol with Message Security Protocol Message Format Specification*. http://www.armadillo.huntsville.al.us/Fortezza_docs/obsolete.html, January 30 1995.
- [4] *FORTEZZA Certification Requirements for Email Applications*. <http://www.armadillo.huntsville.al.us>, 12 February 1996.
- [5] National Security Agency and SDNS Vendor Participants. Secure Data Network System SDNS Message Security Protocol (MSP). SDN.701 Revision 3.0, 21 March 1984.
- [6] Milica Barjaktarović. Formal Specification and Verification of MISSI Architecture Using SPIN. Technical Report 4, AFOSR Summer Research Program, Rome Laboratory, Rome NY, September 1996.
- [7] Milica Barjaktarović. Formal Specification and Verification of MISSI Local Cashe Using SPIN. Technical report, Rome Laboratory and ORA, Rome Laboratory, Rome NY, December 1996.
- [8] Gerard J. Holtzman. *Validation and Verification of Communication Protocols*. Prentice Hall, New York, 1989.

- [9] Gerard J. Holzmann. Basic SPIN Manual. AT&T Bell Laboratories, Murray Hill, New Jersey, 079074.
- [10] Gerard J. Holzmann. Design and Validation of Protocols: a Tutorial. *Computer Networks and ISDN Systems*, 25:981-1017, 1993.
- [11] ITU-T. Data Communication Networks Message Handling Systems, Recommendation x.400. Technical report.
- [12] ITU-T. Data Communication Networks Message Handling Systems, Recommendations x.500-x.521. Technical report.
- [13] National Security Agency, <http://www.armadillo.huntsville.al.us>. *Network Security Managers (NSM) Functional Requirements Specification and Concept of Operations (CONOP)*, June 1996.
- [14] National Security Agency, <http://www.armadillo.huntsville.al.us>. *Network Security Managers (NSM) Functional Requirements Specification and Concept of Operations (CONOP)*, June 1996.
- [15] J.M. Troya P. Merino. Modeling and Verification of the ITU-T Multipoint Communication Service with SPIN. In *Proceedings of the 2nd International Workshop on the SPIN Verification System*, DIMACS, Rutgers University, New Brunswick, New Jersey, August 1996.
- [16] Combined Communications Electronic Board with the Allied Message handling International Subject Matter Experts working group. Common Messaging Strategy and Procedures. ACP 123, November 1994.
- [17] Combined Communications Electronic Board with the Allied Message handling International Subject Matter Experts working group. Common Messaging Strategy and Procedures Annex A: Military Messaging Content Type. ACP 123 Annex A, November 1994.

PERFORMANCE SIMULATION OF
RECEIVERS PROCESSING
LOW PROBABILITY OF INTERCEPT SIGNALS

Dr. Daniel Bukofzer, Professor and Chairman
Department of Electrical and Computer Engineering
California State University
Fresno California 93740-8030

Final Report for:

Summer Faculty Research Program
Rome Laboratory

Sponsored by:
Air Force Office of Scientific Research
Bolling Air Force Base, Washington DC
and Rome Laboratory

December 1997

PERFORMANCE SIMULATION OF RECEIVERS PROCESSING LOW PROBABILITY OF INTERCEPT SIGNALS

Dr. Daniel Bukofzer, Professor and Chairman
Department of Electrical and Computer Engineering
California State University
Fresno California 93740-8030

Abstract

In this report, the performance of receivers processing designed LPI (Low Probability of Intercept) signals is described. The LPI signals involve spread spectrum modulation with a common digital communication structure generated by a so-called flip-wave signal. The signal spectrum is spread by direct sequence methods while using various types of wave shapes involving HBE (High Bandwidth Efficiency) pulses imposed on PN (Pseudo Noise) spreading codes. Signals arriving at the receiver are processed (at baseband) by a D&M (Delay and Multiply) receiver that acts as the detector of spread spectrum signals. The effect of sources of interference on the receiver such as noise and channel induced multipath are evaluated. A post D&M receiver processor compares the signal of a squared integral operation on the D&M receiver filtered output with an adjustable threshold level for the purpose of signal presence detection. Performance results via computer system simulations are carried using a block oriented simulation software package known as Signal Processing Workbench (SPW).

Table of contents

| | |
|---|-----------|
| 1 Introduction | 6 |
| 1.1 Direct Sequence Spread Spectrum..... | 6 |
| 1.2 Pseudo Noise Sequence Generator | 6 |
| 1.3 Delay and Multiply Receivers | 6 |
| 1.4 SPW | 7 |
| 2 Transmitter and Receiver System | 8 |
| 2.1 Flip-Wave Generator | 8 |
| 2.2 High Bandwidth Efficiency Pulses | 10 |
| 2.3 PN Generator | 12 |
| 2.4 D&M receiver | 12 |
| 3 Simulations | 13 |
| 3.1 Delay | 13 |
| 3.1.1 <i>Raised cosine pulse</i> | 13 |
| 3.1.2 <i>Rectangular shaped pulse</i> | 15 |
| 3.1.3 <i>Sinc pulse</i> | 16 |
| 3.2 Noise..... | 21 |
| 3.3 Multipath effects | 22 |
| 3.4 Filtering and Integration..... | 24 |
| 4 Conclusions and Recommendations | 27 |
| References | 28 |
| Appendix | 29 |

List of Tables

| | |
|---|----|
| Table 2-1 PN sequence AM | 12 |
| Table 3-1 Spectral lines for Raised Cosine shaped pulses versus delay without CC | 13 |
| Table 3-2 Spectral lines from Raised Cosine versus delay with CC | 14 |
| Table 3-3 Spectral lines from Rectangular shaped pulse versus delay without CC | 15 |
| Table 3-4 Spectral lines from Rectangular shaped pulse versus delay with CC | 16 |
| Table 3-5 SNR for different PN sequences | 21 |
| Table 3-6 Multipath lables with their random chosen delay and scaling factor parameters | 23 |
| Table 3-7 Threshold level extracted from logical 0's and 1's | 25 |

List of Figures

| | |
|---|----|
| Figure 1-1: A PN sequence generator configuration | 6 |
| Figure 1-2: D&M Receiver Block Diagram | 7 |
| Figure 2-1: QPSK modulator block diagram | 8 |
| Figure 2-2: Rectangular Pulse | 10 |
| Figure 2-3: Raised Cosine Pulse | 10 |
| Figure 2-4: Sinc function shaped pulses of duration $T=4T_c$ | 11 |
| Figure 2-5: Sinc function shaped pulses of duration $T=8T_c$ | 11 |
| Figure 2-6: Sinc function shaped pulses of duration $T=16T_c$ | 12 |
| Figure 2-7: Block diagram of the Delay and Multiply Receiver with switch selected CC | 13 |
| Figure 3-1: Magnitude of chip rate frequency spectral for lines Raised Cosine pulses | 15 |
| Figure 3-2: Magnitude of chip rate frequency spectral lines of rectangular shaped pulses | 16 |
| Figure 3-3: Magnitude of chip rate frequency spectral lines for Sinc pulse shaped by Filter A | 17 |
| Figure 3-4: Magnitude of chip rate frequency spectral lines for Sinc pulse shaped by Filter AW .. | 18 |
| Figure 3-5: Magnitude of chip rate frequency spectral lines for Sinc pulse shaped by Filter B | 18 |
| Figure 3-6: Magnitude of chip rate frequency spectral lines for Sinc pulse shaped by Filter BW .. | 19 |
| Figure 3-7: Magnitude of chip rate frequency spectral lines for Sinc pulse shaped by Filter C | 20 |
| Figure 3-8: Magnitude of chip rate frequency spectral lines for Sinc pulse shaped by Filter CW .. | 21 |
| Figure 3-9: Block diagram of multipath with five delayed and scaled signal replicas | 23 |

| | |
|---|----|
| Figure 3-10: Signal detection block diagram system design | 24 |
| Figure 3-11: Used Threshold level for the analyzed systems | 26 |
| Figure 3-12: Separation available for selecting threshold level..... | 26 |
| Figure A-1: Block Diagram Design (BDD): Transmitter and Receiver system | 29 |
| Figure A-2: Block Diagram Design (BDD): The flip-wave generator..... | 29 |
| Figure A-3: Signal Simulation: PN spreading of random data with raised cosine | 30 |
| Figure A-4: Signal Simulation: Flip-wave signal..... | 30 |
| Figure A-5: SAP, Raised Cosine D&M output without CC with variable delay | 31 |
| Figure A-6: SAP, Rectangular shaped pulses D&M output without CC with variable delay | 32 |
| Figure A-7: SAP, Sinc shaped pulses of duration $T=4T_c$ D&M output with variable delay | 33 |
| Figure A-8: SAP, Sinc shaped pulses of duration $T=8T_c$ D&M output with variable delay | 35 |
| Figure A-9: SAP, Sinc shaped pulses of duration $T=16T_c$ D&M output with variable delay..... | 37 |
| Figure A-10: SAP, Noise simulation..... | 39 |
| Figure A-11: SAP, Multipath with one delayed and scaled duplicate | 39 |
| Figure A-12: SAP, Multipath with 5 delayed and scaled duplicates | 39 |
| Figure A-13: Signal Simulation: Threshold detection | 40 |

List of Abbreviations

| | |
|-------|--------------------------------------|
| BDD | Block Diagram Design |
| BPSK | Binary Phase Shift Keying |
| CC | Complex Conjugation |
| D&M | Delay and Multiply |
| DSSS | Direct Sequence Spread Spectrum |
| FWS | Flip-wave Signal |
| HBE | High Bandwidth Efficiency |
| LPI | Low Probability of Intercept |
| OQPSK | Offset Quadrature Phase Shift Keying |
| PN | Pseudo Noise |
| SAP | Simulation Analysis Page |
| SNR | Signal to Noise Ratio |

1 Introduction

1.1 Direct Sequence Spread Spectrum

The so-called spectrum is a frequency-domain representation of the characteristics of a system or signal, showing the operating frequencies or the bandwidth, respectively. Transformation between time-domain and frequency-domain descriptions is carried out by the Fourier transform pair, mathematically defined by,

$$G(f) = \int_{-\infty}^{\infty} g(t) e^{-j2\pi ft} dt \quad g(t) = \int_{-\infty}^{\infty} G(f) e^{j2\pi ft} df \quad (1.1)$$

A spread spectrum system is one in which the transmitted signal is spread over a frequency band that is much wider than the minimum bandwidth required to transmit the information being sent. An example of a spread spectrum system is wideband FM, where the RF spectrum produced is a function of the information bandwidth and the amount of modulation. A type of spread spectrum signaling of interest insofar as this report is concerned is Direct Sequence Spread Spectrum (DSSS) [2], where a carrier signal is modulated by a digital code sequence (which in most cases is a pseudo noise sequence) whose signaling rate is much higher than the information signal bit rate. DSSS systems produce a low level power density spectrum which gives the signal its low probability of intercept (LPI) property. The LPI feature of spread spectrum systems lend themselves to communications and data transmission with message privacy and signal hiding. In the early 1950's these features were mainly of importance to military communications, but nowadays, many commercial applications are to be found in, for example cellular communication systems, position location devices, and multi-user, multi-channel data communication networks [2].

1.2 Pseudo Noise Sequence Generator

PN Sequence generators are feedback shift register systems with a specific register length, producing a digital code sequence of prescribed length and period [2]. The code is made up of binary bits called chips whose repetition rate is a function of the chip rate,

$$R_{rep} = \frac{\text{clock rate in chips per second}}{\text{code length in chips}} \quad (1.2)$$

and the code length is dependent on the number of shift register stages as well as on the feedback connections. Figure 1-1 shows a PN sequence generator configuration with a feedback connection. Specific feedback connections determine the code produced by the sequence generator.

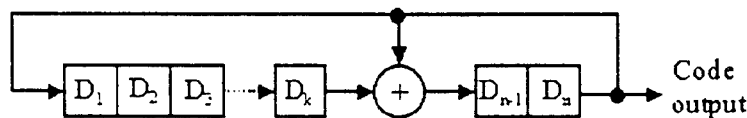


Figure 1-1: A PN sequence generator configuration

1.3 Delay and Multiply Receivers

Delay and Multiply (D&M) receivers are used as presence detectors of a signal in noise and interference. The objective of the detector is to identify the presence of the signal without seeking to determine the actual transmitted message[5]. Known signals are detectable by either passing the received waveform through a filter matched to the signal, or by correlating the received waveform

with a reference that is proportional to the signal. When the signal is not known, matched filters/correlators are not particularly useful because of the required signal knowledge. Digitally modulated signals cannot be considered known unless the sequence that modulates the signal is known. BPSK/DSSS signals are BPSK waveforms [4] with an imposed pseudo noise (PN) spreading sequence. Therefore matched filters/ correlators cannot be used to presence detect BPSK/DSSS signals unless the PN spreading sequence is known. Detection by using a conventional analog spectrum analyzer or FFT processing is difficult because BPSK signals are not periodic (due to the random nature of the sequence) and therefore have continuous Fourier spectra. Furthermore, DSSS signals have a pseudocontinuous spectrum with spectral lines separated in frequency by the reciprocal of the duration of the PN sequence. However when certain nonlinear operations are applied to BPSK for BPSK/DSSS signals, discrete spectral components arise [3]. These components are then often detectable using spectrum analyzer/FFT processing techniques.

Figure 1-2 shows a D&M receiver Block Diagram, where the implementation of a quadratic nonlinearity with which the receiver generates the discrete spectral components at the DSSS signal's chip rate frequency can be seen. Optimization of the signal detectability of the D&M receiver, in the sense of maximizing the spectral SNR of the signal's chip rate frequency spectral line can be estimated by the delay setting "d" of the receiver. A special case corresponding to $d=0$ gives the quadratic structure of the BPSK/DSSS signal. Autocorrelation functions show that a maximum power density at the D&M output is obtained at the chip rate frequency for a delay setting equal to half of the chip duration.

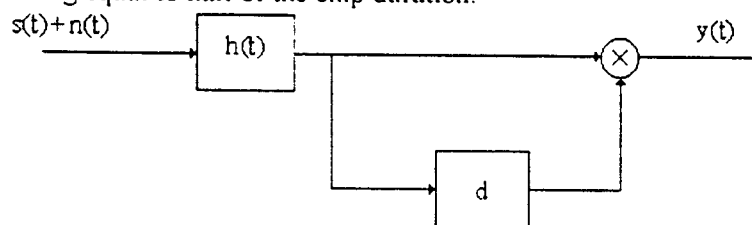


Figure 1-2: D&M Receiver Block Diagram

1.4 SPW

SPW is a UNIX based software package used for the purpose of simulating communication and signal processing systems. SPW has been proved to be a powerful tool to evaluate system performances. The block-orientated characteristic allows building and simulating of complicated systems with the same ease as would be necessary for more simple systems. Schematics are entered in the Block Diagram Editor (BDE) where systems are build up by a hierarchy of blocks from existing libraries. User defined blocks can be saved in a personal library as part of a complete system and as symbols representing the complete system. These blocks and symbols again can be imported into the hierarchy of blocks of some other system. Once a block is added to a schematic, modifications within that block are only effective within that schematic, which gives rise to the possibility of building up identical systems, however working under different conditions. After building a system, simulations can be carried out, while showing the signals of interest in a signal calculator mode. Display of signals of interest following these simulations can be done by

utilization of a built in calculator or a more complicated process such as an FFT to yield the particular analysis of interest.

2 Transmitter and Receiver System

Figure A-1 shows the block diagram generated by the BDE of the system that is used for spreading an input binary random signal with PN sequences shaped by HBE pulses (see section 2.2) and signal detection by a complex delay and multiply receiver. Simulations of the signal spreading operation with PN sequences shaped by raised cosine pulses are shown in Figure A-3. For purposes of this study, the signal is to be detected by a D&M receiver and a post-processing threshold level detector. A binary random data signal with a specific form of spread spectrum modulation is used as the digital communication signal of interest. The spectrum of this spectrally spread random data signal when processed by a D&M receiver tends to show a spectral line at the chip rate frequency making it therefore somewhat easy to detect. The purpose of this project is to evaluate the detectability of this signal under various operational scenarios involving noise and multipath effects on the signal and receiver processing modifications. In order to reduce the detectability of this signal, its spectrum before transmission has to appear featureless. A featureless signal will show a spectrum that to the observer appears like unusable noise without any significant spectral lines. The flip-wave signal generator is a device by which after spectrum spreading, a featureless signal as described below is produced.

2.1 Flip-Wave Generator

The basis of the flip-wave signal involves QPSK (Quadrature Phase Shift Keying) modulation. A conventional QPSK modulator is shown in Figure 2-1.

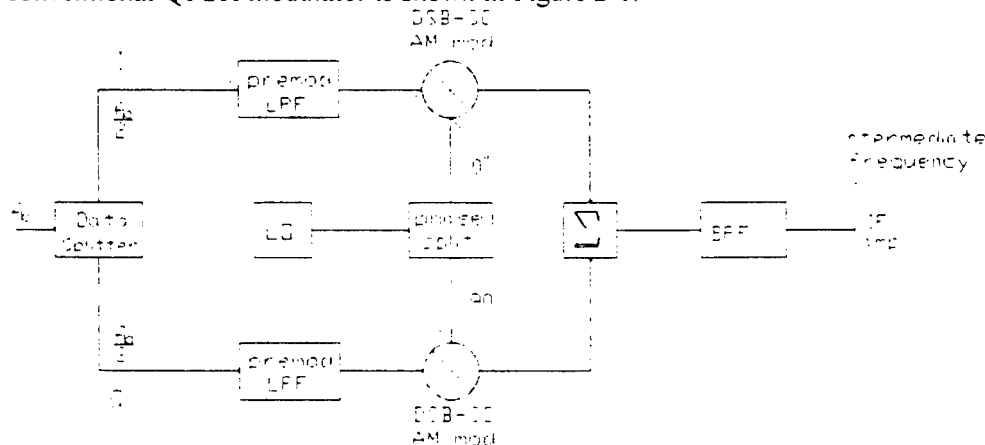


Figure 2-1: QPSK modulator block diagram

The random data input signal is mathematically described as

$$Z(t) = \sum_{k=-\infty}^{\infty} A_k p(t - kT_b) \quad -\infty < t < \infty \quad (2.1)$$

where the data $\{A_k\}$ is a ± 1 valued random sequence and $p(t)$ is a deterministic pulse.

It is assumed that

$$E\{A_k\} = 0 \quad \forall k \quad (\text{i.e., r.v.'s are zero mean}) \quad (2.2)$$

$$E\{A_k A_n\} = \begin{cases} 1 & k \neq n \\ 0 & k = n \end{cases} \quad (\text{i.e., r.v.'s are independent and identically distributed}) \quad (2.3)$$

and

$$p(t) = \begin{cases} 1 & 0 < t < T_b \\ 0 & \text{otherwise} \end{cases} \quad (2.4)$$

After splitting $Z(t)$ into two components, the "In phase" baseband signal is described as

$$Z_I(t) = \sum_{k=-\infty}^{\infty} A_{2k} p'(t - 2kT_b) = \sum_{k=-\infty}^{\infty} A_{2k} p\left(\frac{t}{2} - kT_b\right) \quad (2.5)$$

and the "Quadrature" baseband signal as

$$Z_Q(t) = \sum_{k=-\infty}^{\infty} A_{2k+1} p'(t - (2k+1)T_b) = \sum_{k=-\infty}^{\infty} A_{2k+1} p\left(\frac{t}{2} - (k + \frac{1}{2})T_b\right) \quad (2.6)$$

where

$$p'(t) \triangleq p(t/2) \quad (2.7)$$

A so-called offset QPSK (OPQSK) modulated signal results when amplitude modulation of sine and cosine carriers is imposed. That is,

$$\begin{aligned} S_{OPQSK}(t) &= A[Z_I(t) \cos 2\pi f_c t - Z_Q(t) \sin 2\pi f_c t] \\ &= A \cos[2\pi f_c t + \phi(t)] \end{aligned} \quad (2.8)$$

The carrier phase now is

$$\phi(t) = \tan^{-1} \frac{Z_Q(t)}{Z_I(t)} \quad (2.9)$$

and can take on values $\pm\pi/4, \pm3\pi/4$ as a result of the fact that $Z_I(t)$ and $Z_Q(t)$ are bipolar waveforms. The waveform produced by the flip-wave generator is built up according to the following algorithm: when $A_k=1$ the carrier phase in the interval $(k+1)T_b < t < (k+2)T_b$ is that of the carrier in the interval $kT_b < t < (k+1)T_b$ incremented by $\pi/2$ radians, and conversely when $A_k=-1$ the carrier is incremented by $-\pi/2$ radians. The BDE generated block diagram of the flip-wave generator is shown in Figure A-2 of the Appendix. The flip-wave generator used to produce the LPI signal generates a waveform by defining two sequences of r.v.'s, $\{X_k\}$ and $\{Y_k\}$ which are related to $\{A_k\}$ by the recursions

$$X_{k+1} = -A_k Y_k \quad Y_{k+1} = A_k X_k \quad k=0, 1, 2, \dots \quad (2.10)$$

The initial values

$$X_0 = 1 \quad Y_0 = 0 \quad (2.11)$$

define the starting time $t=0$. Now $X(t)$ and $Y(t)$ are defined by

$$X(t) = \sum_{k=0}^{\infty} X_k p(t - kT_b) \quad Y(t) = \sum_{k=0}^{\infty} Y_k p(t - kT_b) \quad (2.12)$$

In Figure A-4, the signal simulation of the one-input/two-output flip-wave generator and the transformation signals specified by Equation 2.10 and Equation 2.12 with the random data input are shown. These waveforms take on the values 0, ± 1 in such way that if $X(t)$ equals ± 1 , then $Y(t)$

must be zero valued and vice versa. This is the basis of the so-called flip-wave signal which is an example of differential phase modulation in an OQPSK format where at any one signaling interval, a phase change of $\pm\pi/2$ rad. must take place. DSSS modulation for spreading of the flip-wave signal is accomplished by generating spreading codes, using linear feedback shift register systems and various shaped HBE pulses. Detailed mathematical descriptions [1] of these spreading methods results in the expression of the spread flip-wave signal processed by the receiver, to be given by

$$S'_{FW\Sigma}(t) = \sum_{k=0}^n X_{2k} p'(t - 2kT_b) \sum_{m=-\infty}^n \alpha_m h(t - mT) + j \sum_{k=0}^n Y_{2k+1} p'(t - (2k+1)T_b) \sum_{l=0}^n \beta_l h(t - (2l+1)T/2) \quad (2.13)$$

where $\{\alpha_m\}$ and $\{\beta_l\}$ are ± 1 valued M-sequences, T^{-1} is the chip rate of the spreading codes, and $h(t)$ is one of the various HBE pulses imposed on the PN spreading codes.

2.2 High Bandwidth Efficiency Pulses

The various HBE pulses considered are described as follows:

1. Rectangular shaped pulses

Let T be the duration of one pulse period, and the rectangular pulse be described by

$$h(t) = \begin{cases} 1 & 0 < t < T = 2T_c \\ 0 & \text{otherwise} \end{cases} \quad (2.14)$$

where the time shift between spreading codes is

given by, $\frac{T}{2} = T_c$. (T^{-1} is the chip rate as indicated above.)

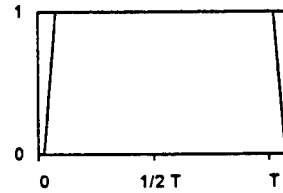


Figure 2-2: Rectangular Pulse

The HBE pulses shapes to follow are normalized in amplitude to that of the rectangular pulse shape in order to obtain equal signal energies.

2. Raised Cosine shaped pulses

$$h(t) = \begin{cases} \sin^2(2\pi t/4T) & 0 < t < T = 2T_c \\ 0 & \text{otherwise} \end{cases} \quad (2.15)$$

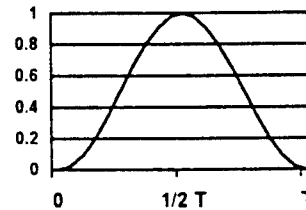


Figure 2-3: Raised Cosine Pulse

I. Sinc shaped pulses

The implementation of the sinc shaped pulses is obtained by exciting finite impulse response (FIR) filters where the filter coefficients are set to correspond to the numerical amplitude values of the pulse function. Limited numbers of coefficients in each filter block are the reason why for longer duration pulses parallel filters with appropriate time delays are used in order to accomplish the desired implementation.

A. Sinc function shaped pulses of duration $T=4T_c$ realized using what is labeled as Filter

A.

$$h_{SINC}(t) = \frac{\sin(2\pi(t - 2T_c)/4T_c)}{2\pi(t - 2T_c)/4T_c} \quad 0 < t < T \quad (2.16)$$

b) Hanning weighted sinc function shaped pulses of duration $T=4T_c$ realized using what is labeled as Filter AW.

$$h_{HWSINC}(t) = h_{SINC}(t) * \sin^2\left(\frac{\pi t}{4T_c}\right) \quad 0 < t < T \quad (2.17)$$

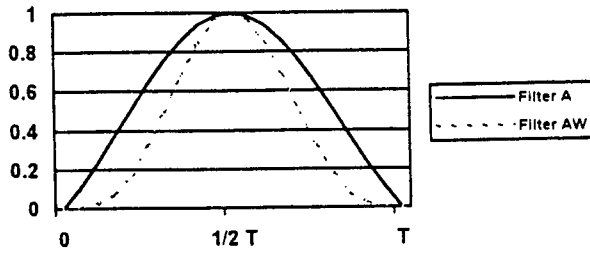


Figure 2-4: Sinc function shaped pulses of duration $T=4T_c$.

- c) Sinc function shaped pulses of duration $T=8T_c$ realized using what is labeled as Filter B,

$$h_{SINC}(t) = \frac{\sin(2\pi(t-4T_c)/4T_c)}{2\pi(t-4T_c)/4T_c} \quad 0 < t < T \quad (2.18)$$

- d) Hanning weighted sinc function shaped pulses of duration $T=8T_c$ realized using what is labeled as Filter BW.

$$h_{HWSINC}(t) = h_{SINC}(t) * \sin^2\left(\frac{\pi t}{8T_c}\right) \quad 0 < t < T \quad (2.19)$$

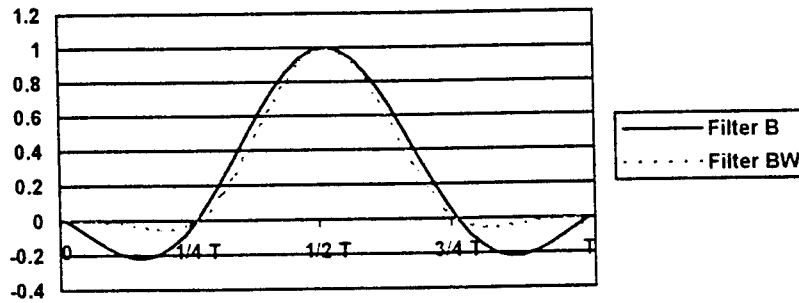


Figure 2-5: Sinc function shaped pulses of duration $T=8T_c$.

- e) Sinc function shaped pulses of duration $T=16T_c$ realized using what is labeled as Filter C,

$$h_{SINC}(t) = \frac{\sin(2\pi(t-8T_c)/4T_c)}{2\pi(t-8T_c)/4T_c} \quad 0 < t < T \quad (2.20)$$

- f) Hanning weighted sinc function shaped pulses of duration $T=16T_c$ realized using what is labeled as Filter CW.

$$h_{HWSINC}(t) = h_{SINC}(t) * \sin^2\left(\frac{\pi t}{16T_c}\right) \quad 0 < t < T \quad (2.21)$$

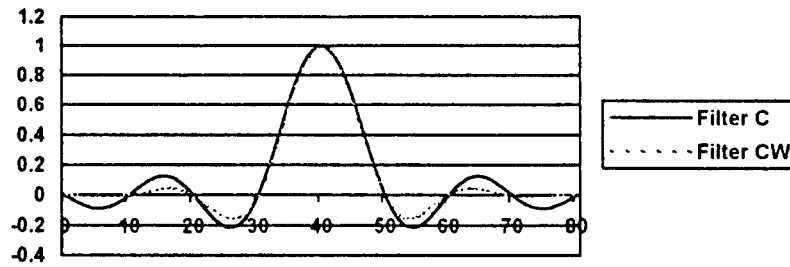


Figure 2-6: Sinc function shaped pulses of duration $T=16T_c$

2.3 PN Generator

Three parallel m-sequence PN generators are used to produce amplitude modulation (AM) on the chips. The number of AM levels depends on the weight assigned to each parallel generator; hence the possible outcomes of the amplitude variable α_m , with its corresponding probabilities. Normalization of the different amplitudes is realized by equalizing the output by the corresponding mean squared values, where

$$E\{\alpha_m^2\} = \sum_{n=1}^m \alpha_n^2 \cdot P(\alpha_n) \quad (2.22)$$

The different cases are summarized in Table 2-1.

| chip weight | # of levels | α_m | probability | $E\{\alpha_m^2\}$ |
|-------------|-------------|--|--|-------------------|
| 1.1.1 | 4 | ± 3 ± 1 | $\frac{1}{4}$ $\frac{3}{4}$ | 3 |
| 4.1.1 | 6 | ± 6 ± 4 ± 2 | $\frac{1}{4}$ $\frac{1}{2}$ $\frac{1}{4}$ | 18 |
| 4.2.1 | 8 | ± 7 ± 5 ± 3 ± 1 | $\frac{1}{4}$ $\frac{1}{4}$ $\frac{1}{4}$ $\frac{1}{4}$ | 21 |

Table 2-1 PN sequence AM

2.4 D&M receiver

The complex D&M receiver can operate in two different switch selected modes. There is a choice of processing the baseband signal with or without complex conjugation prior to multiplication. If complex conjugation is selected, the input signal is multiplied by its complex conjugated delayed version. This is shown in Figure 2-7.

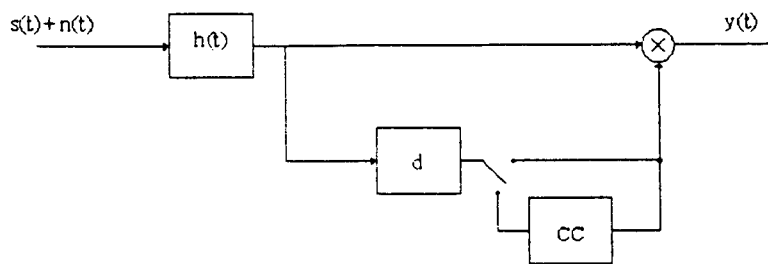


Figure 2-7: Block diagram of the Delay and Multiply Receiver with switch selected CC

3 Simulations

3.1 Delay

One of the variables that affects the output of the D&M receiver is its delay setting. By changing the delay of the D&M receiver, the spectrum of the receiver output will be affected, making the delay setting an important variable which influences the signal detection probability performance of the receiver. For different types of HBE pulses the effect of the delay setting on the performance of the D&M receiver as well as the effect of using complex conjugation prior to multiplication is analyzed.

3.1.1 Raised cosine pulse

For simulation purposes a raised cosine pulse is used with a chip rate frequency of $f_c = 256$ kHz. The time shift between spreading codes is given by,

$$T_c = \frac{T}{2} = 1.9531 \mu\text{sec} \quad (3.1)$$

With a simulation sampling frequency set to $f_s = 5.120$ kHz, one chip interval represents $f_s / f_c = 20$ samples. The presence of spectral lines in the spectrum at the output of the D&M receiver under different delay settings for raised cosine shaped pulses are described and summarized in Table 3-1 and Table 3-2. Figure A-5 in the Appendix shows some of the spectra obtained in the SPW signal calculator mode for raised cosine shaped pulses with and without complex conjugation. It is observed that,

| Raised cosine pulse without Complex Conjugation | |
|---|--|
| Delay (sample points) | visible spectral lines |
| Delay=0 | f_c phase of spectrum is zero for $4 f_c, 6 f_c, 8 f_c, 10 f_c$ |
| Delay=1-2 | f_c zero-phases disappear |
| Delay=3-4 | $f_c, 5 f_c, 7 f_c$ |
| Delay=5 | $f_c, 3 f_c, 5 f_c, 9 f_c$ |
| Delay=6 | $f_c, 3 f_c, 5 f_c, 7 f_c$ |
| Delay=7-10 | $f_c, 3 f_c, 5 f_c, 7 f_c, 9 f_c$ |
| Delay=11-15 | $3 f_c, 5 f_c, 7 f_c$ |
| Delay=16 | $5 f_c, 7 f_c$ |
| Delay=17-20 | No significant spectral lines |

Table 3-1 Spectral lines for Raised Cosine shaped pulses versus delay without CC

- Using delay settings of less than or equal to half a chip duration results in the presence of strong spectral lines at the frequency f_c in the output of the D&M receiver.
- Spectral lines at frequencies which are multiples of the chip rate frequency do not exceed the spectral power level around the chip rate frequency.
- The spectral line at the frequency f_c at the output of the D&M receiver without Complex Conjugation disappears for delays greater than half a chip duration (10 sample delay). Also in this case the spectral lines at frequencies which are multiples of f_c do not exceed the spectral power level around the chip rate frequency.
- At a receiver delay setting near 1 chip duration (20 samples) in length, all spectral lines at the output disappear.

| Raised cosine pulse with Complex Conjugation | |
|--|--|
| Delay (sample points) | visible spectral lines |
| Delay=0-2 | 0Hz, $2 f_c$ |
| Delay=3 | 0Hz, $2 f_c$, $6 f_c$ |
| Delay=4 | 0Hz, $2 f_c$, $4 f_c$, $6 f_c$ |
| Delay=5-9 | 0Hz, $2 f_c$, $4 f_c$, $6 f_c$, $8 f_c$ |
| Delay=10 | 0Hz, $2 f_c$ |
| Delay=11 | 0Hz, $2 f_c$, $4 f_c$, $6 f_c$, $8 f_c$ |
| Delay=12-13 | $2 f_c$, $4 f_c$, $6 f_c$, $8 f_c$ |
| Delay=14 | $4 f_c$, $6 f_c$, $8 f_c$ |
| Delay=15-16 | $4 f_c$, $6 f_c$ |
| Delay=17-20 | No significant spectral lines |

Table 3-2 Spectral lines from Raised Cosine versus delay with CC

- Receivers with CC produce at their output spectral lines at even valued multiples of f_c versus odd valued multiples of f_c for receivers without CC.
- A further difference between a receiver using Complex Conjugation and a receiver that does not, is that the former produces an output spectral line at zero frequency, as long as it is not suppressed due to the use of a large value of delay.

Spectral Strength at the Chip Rate Frequency with Variable Receiver Delay for Raised Cosine shaped pulses

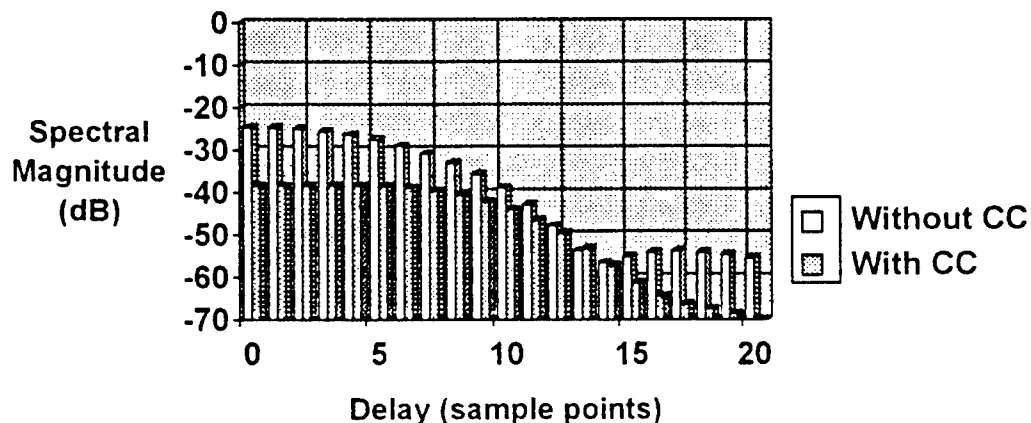


Figure 3-1: Magnitude of chip rate frequency spectral for lines Raised Cosine pulses

- The main lobe bandwidth of the receiver output spectrum is approximately $8 \cdot f_c$.
- The spectral magnitude of the chip rate frequency component is maximum for small delays.
- A spectral strength of approximately -50 dB is considered to be the level at which the spectral lines are not detectable when compared to the rest of the spectrum.
- For delays greater than 10 samples (half a chip duration interval) the spectral line at the chip rate frequency cannot be discriminated from the rest of the spectrum.

Figure 3-1 shows the changing strength of the spectral component at f_c as a function of delay for a receiver that uses or bypasses complex conjugation.

3.1.2 Rectangular shaped pulse

The same chip rate frequency of $f_c=256\text{kHz}$ and sampling frequency of $f_s=5.120\text{ksps}$ are used as in the case of Raised Cosine pulses. Again, one chip duration interval represents $f_s/f_c=20$ samples. The presence of spectral lines in the spectrum at the output of the D&M receiver under different delay settings for rectangular shaped pulses are described and summarized in Table 3-3 and Table 3-4. Figure A-6 in the Appendix shows some of the spectra obtained in the SPW signal calculator mode for rectangular shaped pulses. It is observed that.

| Rectangular shaped pulse without Complex Conjugation | |
|--|-----------------------------------|
| Delay (sample points) | visible spectral lines |
| Delay=0 | No significant spectral lines |
| Delay=1 | $3 f_c, 5 f_c, 7 f_c, 9 f_c$ |
| Delay=2 | $f_c, 3 f_c, 5 f_c, 7 f_c, 9 f_c$ |
| Delay=3 | $f_c, 3 f_c, 5 f_c, 9 f_c$ |
| Delay=4-10 | $f_c, 3 f_c, 5 f_c, 7 f_c, 9 f_c$ |
| Delay=11 | $f_c, 3 f_c, 5 f_c, 7 f_c$ |
| Delay=12 | $f_c, 3 f_c, 7 f_c, 9 f_c$ |
| Delay=13 | $f_c, 5 f_c, 7 f_c, 9 f_c$ |
| Delay=14-15 | $f_c, 3 f_c, 5 f_c, 7 f_c, 9 f_c$ |
| Delay=16 | $f_c, 3 f_c, 7 f_c, 9 f_c$ |
| Delay=17 | $f_c, 3 f_c, 5 f_c, 9 f_c$ |
| Delay=18 | $f_c, 3 f_c, 5 f_c, 7 f_c, 9 f_c$ |
| Delay=19 | $3 f_c, 5 f_c, 7 f_c, 9 f_c$ |
| Delay=20 | No significant spectral lines |

Table 3-3 Spectral lines from Rectangular shaped pulse versus delay without CC

- Spectral lines at the frequency f_c are strong for receiver delays set between approximately 4 and 17 sample points.
 - Spectral lines at frequencies which are multiples of the chip rate frequency do not exceed the spectral power level around the chip rate frequency.
- At a receiver delay set to 1 chip duration (20 samples) all spectral lines at the output disappear.

| Rectangular shaped pulse with Complex Conjugation | |
|---|--|
| Delay (sample points) | visible spectral lines |
| Delay=0 | 0Hz |
| Delay=1-4 | 0Hz, 2 f_c , 4 f_c , 6 f_c , 8 f_c |
| Delay=5 | 0Hz, 2 f_c , 6 f_c |
| Delay=6-9 | 0Hz, 2 f_c , 4 f_c , 6 f_c , 8 f_c |
| Delay=10 | 0Hz |
| Delay=11-14 | 0Hz, 2 f_c , 4 f_c , 6 f_c , 8 f_c |
| Delay=15 | 0Hz, 2 f_c , 6 f_c |
| Delay=16-17 | 0Hz, 2 f_c , 4 f_c , 6 f_c , 8 f_c |
| Delay=18-19 | 2 f_c , 4 f_c , 6 f_c , 8 f_c |
| Delay=20 | No significant spectral lines |

Table 3-4 Spectral lines from Rectangular shaped pulse versus delay with CC

- The receiver with CC produces at its output spectral lines at even valued multiples of f_c versus odd valued multiples of f_c for the receiver without CC.
- The spectral line at zero frequency, if not suppressed due to high receiver delay settings is visible at the output.

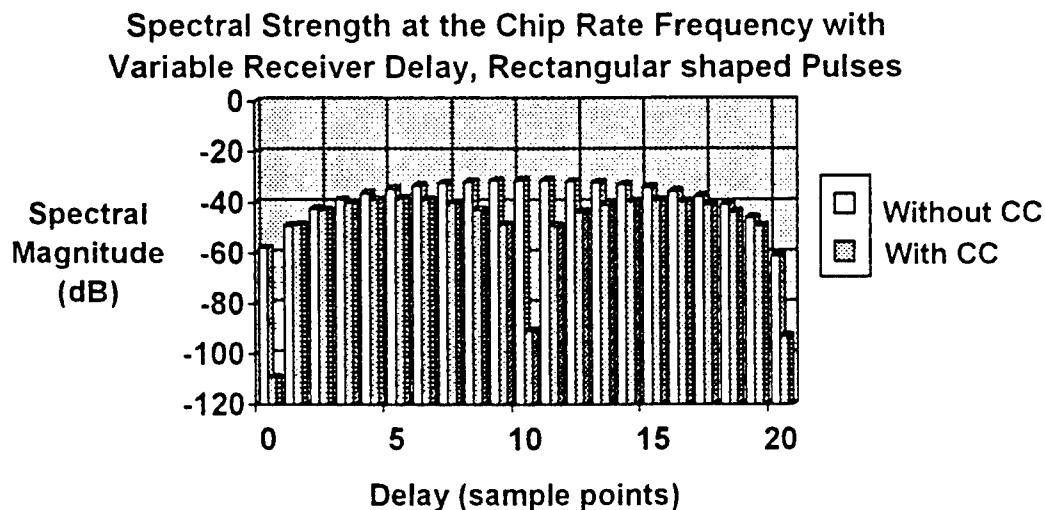


Figure 3-2: Magnitude of chip rate frequency spectral lines of rectangular shaped pulses

- The main lobe bandwidth of the output spectrum varies between approximately $2*f_c$ and $8*f_c$ Hz for the different receiver delay settings.
- A spectral strength of approximately -65 dB is considered to be the level at which the spectral lines are not detectable when compared to the rest of the spectrum.

Figure 3-2 shows the changing strength of the spectral component at f_c as a function of delay for a receiver that uses or bypasses complex conjugation.

3.1.3 Sinc pulse

- Analogous to the above described simulations, the strength of the chip rate frequency spectral lines are presented graphically in Figure 3-3 to Figure 3-8 for cases in which the pulse shape

are of various sinc function form as described in Section 2.2. One chip duration is equal to 20 samples. The receiver output spectra observed from the various simulations carried out are shown in Figures A-7 to A-9 in the Appendix. The focus of the output spectral lines for systems with CC is at frequencies that are twice as high (512 kHz) as for systems without CC (256 kHz).

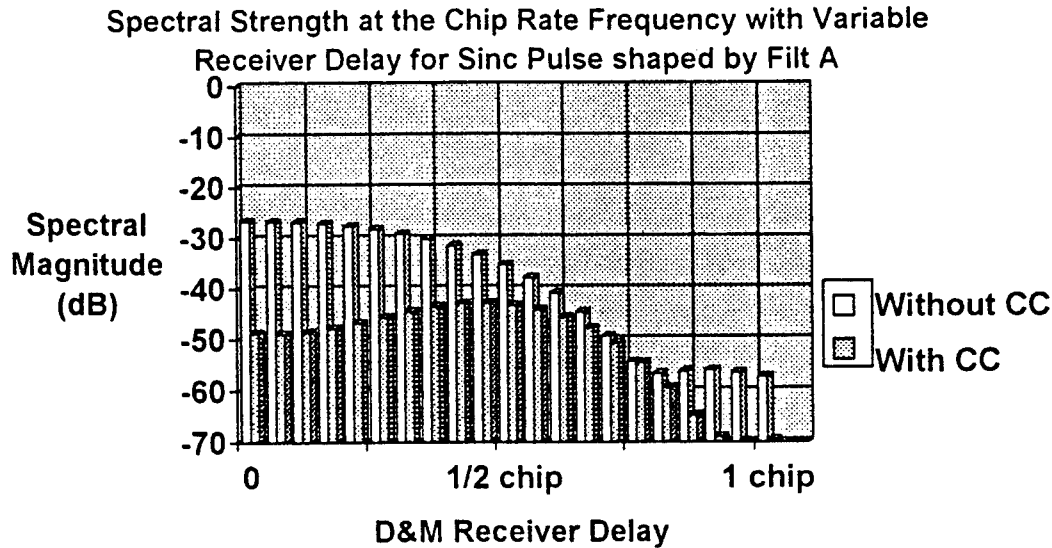


Figure 3-3: Magnitude of chip rate frequency spectral lines for Sinc pulse shaped by Filter A

Simulation results for D&M receivers processing a spread flip-wave signal with sinc shaped HBE pulses described by Equation (2.16) are shown in Figure A-7a and Figure A-7b. Referring to these simulation results the following observations can be made.

- The main lobe bandwidth of the output spectrum varies between approximately $5 \cdot f_c$ and $7 \cdot f_c$.
- The spectral line at the chip rate frequency for both, a receiver using or bypassing complex conjugation is not visible anymore for delays greater than 14 samples, where 20 samples represent the length T .
- A spectral strength of approximately -50 dB is considered to be the level at which the spectral lines are not detectable when compared to the rest of the spectrum.
- The spectral lines at the output of the system with CC for delays smaller than the mentioned 14 samples are all clearly visible, but the strength may not always exceed the set -50 dB minimum level especially for the case of smaller delays.

Figure 3-3 shows the changing strength of the spectral component at f_c as a function of delay for a receiver that uses or bypasses complex conjugation.

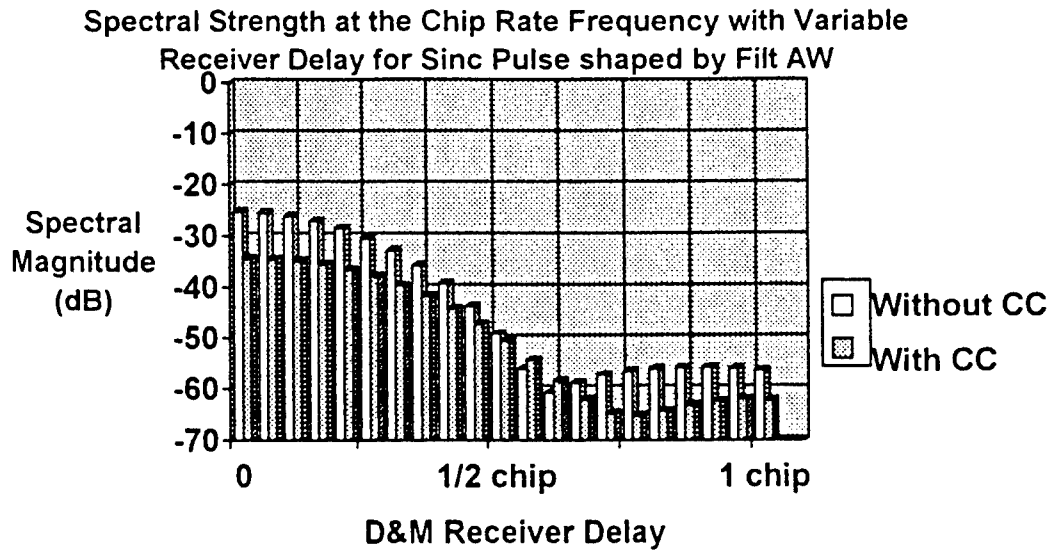


Figure 3-4: Magnitude of chip rate frequency spectral lines for Sinc pulse shaped by Filter AW

Similar observations were made for simulations with D&M receivers processing a spread flip-wave signal with sinc shaped HBE pulses described by Equation (2.17). These results are shown in Figure A-7c and Figure A-7d as was done when HBE pulses from Equation (2.16) were used. The only difference here is that the main lobe bandwidth of the output spectrum varies between approximately $8 \cdot f_c$ and $10 \cdot f_c$. Figure 3-4 shows the changing strength of the spectral component at f_c as a function of delay for a receiver that uses or bypasses complex conjugation.

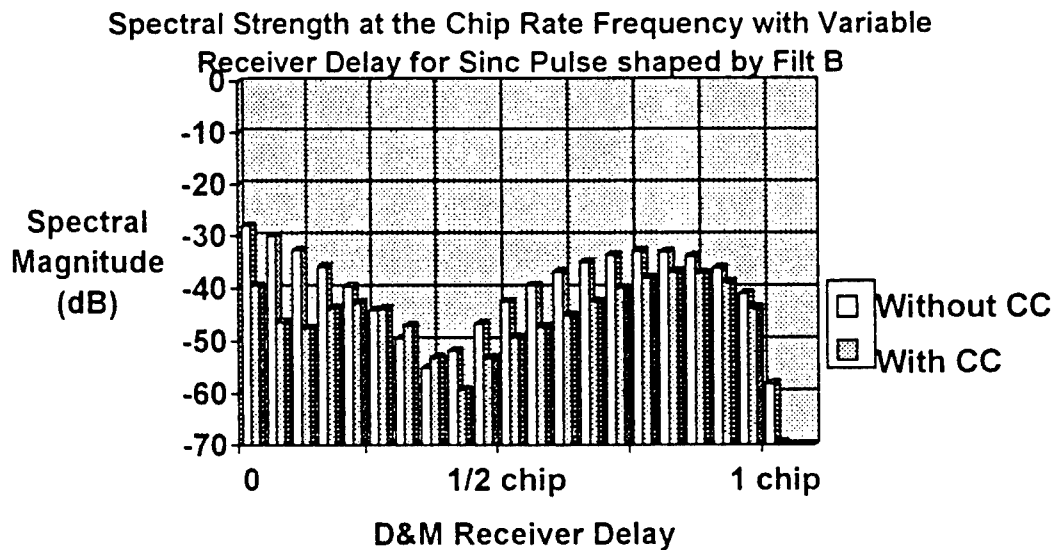


Figure 3-5: Magnitude of chip rate frequency spectral lines for Sinc pulse shaped by Filter B

Figure A-8a and Figure A-8b show simulation results for D&M receivers processing a spread flip-wave signal with sinc shaped HBE pulses described by Equation (2.18). Observations made for simulations of D&M receivers processing a spread flip-wave signal with sinc shaped HBE pulses described by Equation (2.18) are.

- A main lobe of the output spectrum is not clearly visible.
- For receiver delays around 6 to 8 samples (20 samples are equal to 1 chip duration interval) the spectral line of the chip rate frequency does not exceed the spectral strength of surrounding spectral lobes.
- A spectral magnitude between -45dB and -50 dB is considered to be the level at which the spectral lines are not detectable when compared to the rest of the spectrum.
- The spectral lines at the output of the system with CC for delays smaller than the mentioned 14 samples are all clearly visible, but the strength may not always exceed the set -50 dB minimum level especially for the case of smaller delays.

Figure 3-5 shows the changing strength of the spectral component at f_c as a function of delay for a receiver that uses or bypasses complex conjugation.

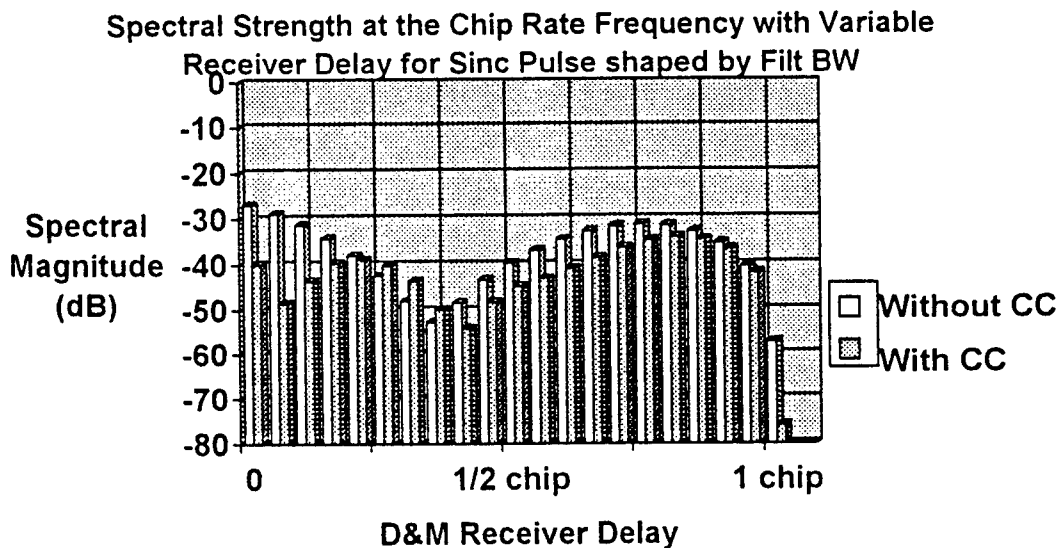


Figure 3-6: Magnitude of chip rate frequency spectral lines for Sinc pulse shaped by Filter BW

Figure A-8c and Figure A-8d support the following observations. Difference between D&M receivers processing a spread flip-wave signal with sinc shaped HBE pulses described by Equation (2.18) and receivers processing the Hanning weighted sinc shaped pulses from Equation (2.19) is only observable in small spectral strength differences at the frequencies of interest. Figure 3-6 shows the changing strength of the spectral component at f_c as a function of delay for a receiver that uses or bypasses complex conjugation.

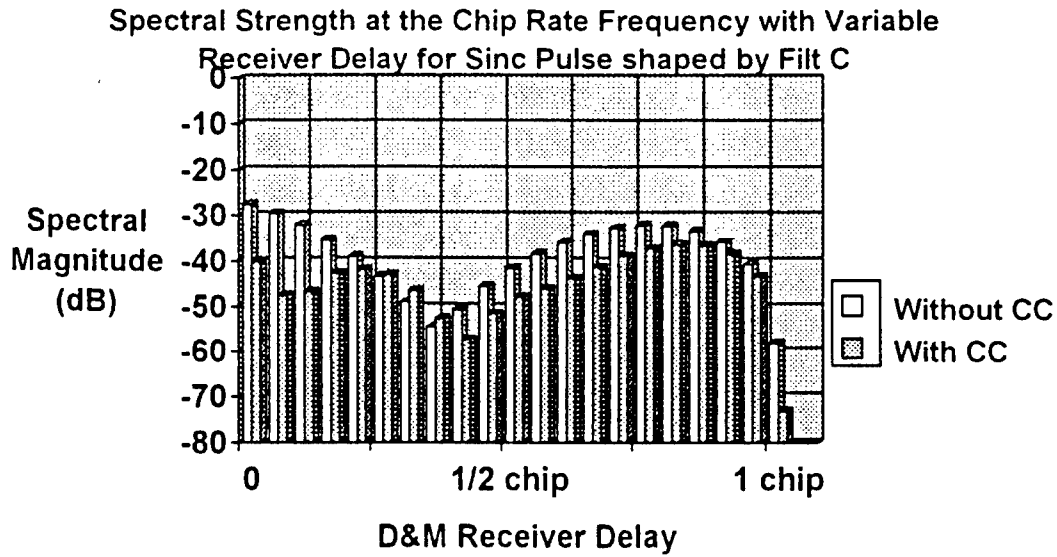


Figure 3-7: Magnitude of chip rate frequency spectral lines for Sinc pulse shaped by Filter C. Simulation results displayed in Figure A-9a and Figure A-9b show the case of D&M receivers processing a spread flip-wave signal with sinc shaped HBE pulses described by Equation (2.20). The following observations can be made.

- A main lobe of the output spectrum is not clearly visible.
- For receiver delays around 6 to 8 samples (20 samples is equal to T in time duration) the spectral line of the chip rate frequency does not exceed the spectral strength of surrounding spectral lobes.
- Due to different strengths of the spectral lobes, a spectral strength between approximately -45dB and -50 dB is considered to be the level at which the spectral lines are not detectable when compared to the rest of the spectrum.
- The spectral lines at the output of the system with CC for delays smaller than the mentioned 14 samples are all clearly visible, but the strength may not always exceed the set -50 dB minimum level especially for the case of smaller delays.

Figure 3-7 shows the changing strength of the spectral component at f_c as a function of delay for a receiver that uses or bypasses complex conjugation.

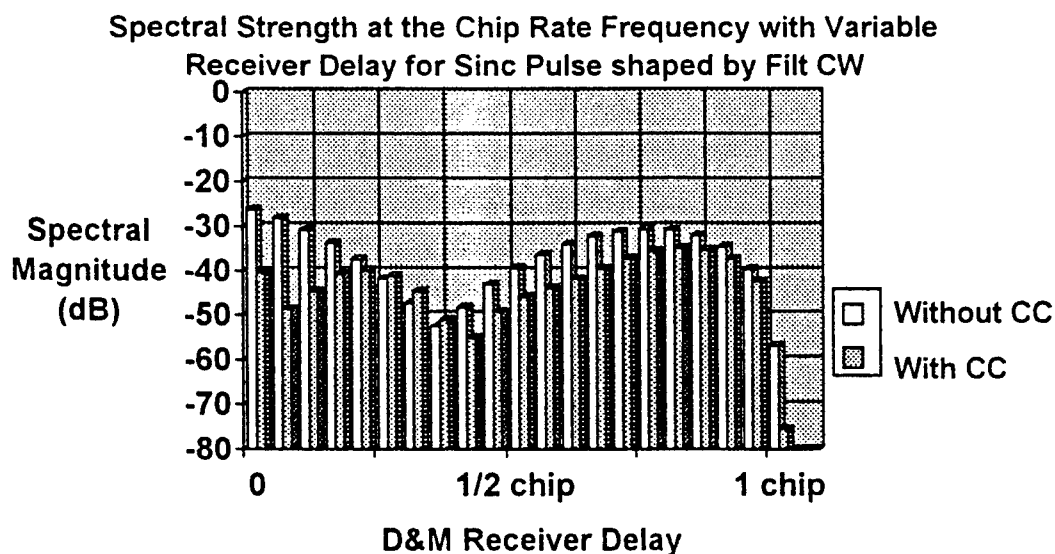


Figure 3-8: Magnitude of chip rate frequency spectral lines for Sinc pulse shaped by Filter CW. When simulations are run for D&M receivers processing a spread flip-wave signal with sinc shaped HBE pulses described by Equation (2.21), similar observations can be made as for receivers processing sinc shaped HBE pulses described by Equation (2.20). The results are shown in Figure A-9c and Figure A-9d. Figure 3-8 shows the changing strength of the spectral component at f_c as a function of delay for a receiver that uses or bypasses complex conjugation.

3.2 Noise

In practical situations, a signal is transmitted over a channel where the signal will be susceptible to noise interference. Noise in the channel and receiver system will affect the performance of the D&M receiver. For modeling purposes, the noise is added to the signal of interest before it enters the D&M receiver. The noise-level and its bandwidth can be set in the simulation of the system. Normalization of the AM modulator with three parallel scaled m-sequence PN generators insures that the possible AM signal levels have a minimum set distance from each other. To distinguish the different levels of the AM modulator no more noise is added to the signal than specified by the minimum signal to noise ratios shown in Table 3-5 for the different cases of chip weight. Although this is not necessarily the case in practical systems, the approach used insures that the signal is not buried in the noise. (See also Table 2-1.) Therefore the added system noise is limited such that

$$SNR = -20 \log \left[E \{ \alpha_m^2 \} \right] \quad (3.2)$$

| # of levels | $E \{ \alpha_m^2 \}$ | SNR (dB) |
|-------------|----------------------|----------|
| 4 | 3 | 9 |
| 6 | 18 | 25 |
| 8 | 21 | 26 |

Table 3-5 SNR for different PN sequences

Simulations are run for a receiver operating SNR with values specified in Table 3-5. Systems without any noise have $SNR = \infty$. The selected noise bandwidth is as wide as the signal bandwidth. Doubling of the noise bandwidth results in a noise amplitude spectral strength multiplied by a factor $\sqrt{2}$ in the simulations with SPW.

Random white noise has a flat frequency spectrum, where all frequencies are present in equal strength. Therefore such noise is clearly featureless. Adding noise to the signal results in the strength (in dB) of the system output spectrum to be raised equally over the entire frequency range. The lower the SNR, the more the output spectrum characteristics become that of a featureless signal. However the spectral line component at the chip rate frequency, always remains visible due to the limitations placed on the SNR. Figure A-10 of the Appendix shows the output spectrum of a D&M receiver processing a signal without noise, the spectrum of pure noise and the output spectrum of a D&M receiver processing a signal with a minimum SNR.

3.3 Multipath effects

A major source of interference in communication applications is channel induced multipath which results in the received signal to be made up of the sum of scaled and delayed versions of the transmitted signal. The next set of observations relate to D&M receiver simulation results with one level of multipath. That is, the original signal together with one scaled and delayed replica arrive at the receiver. Figure A-11 shows some D&M receiver output spectra after processing signals with various one level multipath delays and magnitudes.

- When the scaled and delayed component due to multipath has a delay smaller or equal to half a chip duration interval and a magnitude as strong as the transmitted signal, the effect of multipath is significant on the output spectrum of the D&M receiver as the spectral lines at the chip rate frequency and multiples of it disappear in these cases. In Figure A-11 the D&M receiver output spectrum of a signal with its duplicate due to multipath is shown where the delay is half a chip duration and the duplicated signals' strength is varied.
- Systems with a multipath delay equal to multiples of the chip duration interval (20 samples) show spectral lines at the frequencies of interest in the D&M receiver output spectra that remain unchanged for all multipath strengths that are weaker (or equally as strong) as the transmitted signal.
- Values of the multipath delays from a few chip duration intervals and higher do not cause as much spectral line weakening as for cases of multipath delays with lower values.
- Beyond the characteristic spectral lines, the rest of the spectrum for high values of multipath delay is more flattened and the main lobe of the spectrum is less observable.
- Referring to the different spectral strengths at the chip rate frequency for receivers that do or do not use complex conjugation, it is important to know which type of HBE pulse is used and to which value the delay of the receiver is set to. Namely in the case where HBE rectangular shaped pulses are used, a receiver delay of half a chip duration (10 samples) shows a D&M receiver output with spectral lines that are maximum at the frequencies of interest for receivers that do not use complex conjugation, but which are minimum for receivers that do use CC.
- If the spectrum of the D&M receiver output does not show any spectral lines when no multipath effects are present, there will also not be any in the case when multipath is present.
- To look at the effect of multipath on D&M receivers with complex conjugation, a receiver delay of 11 samples was chosen which resulted in close to maximum spectral lines.

Since multipath is a phenomenon for which the signal scaling and delay parameters are not known to the receiver, random chosen values of scaling and delay were used, in order to simulate a more realistic multipath environment. In the following simulations, multipath is realized containing five scaled and delayed signal replicas. A schematic of the block diagram used to simulate multipath with more than one delayed and scaled signal replica is shown in Figure 3-9.

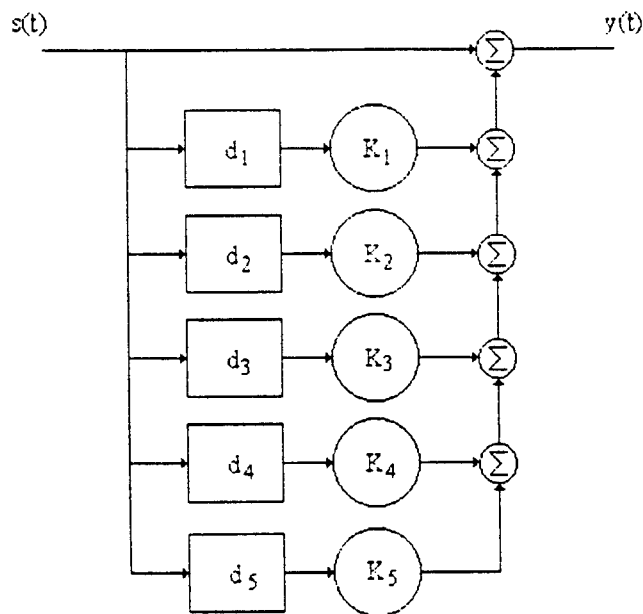


Figure 3-9: Block diagram of multipath with five delayed and scaled signal replicas

For such simulations, random values had to be chosen to set the delay and scaling factor of each signal replica. With a mathematical computer program (MATLAB), uniformly distributed random numbers for the delay and the scaling factor were generated. The delay of each of the five signal components is a random variable with values between 0 and 200 samples, which corresponds to values between 0 and 10 chip duration intervals. The random variable for the scaling factor involves values in a range of 0 and 1.5 times the original signal amplitude. The generated values for the delay and scaling factors are presented in Table 3-6. To compare the different sets of values, the multipath systems are labeled with numbers. For example the multipath system with label number 11 has signal duplicates with delays of 3, 63, 139, 23, 64 samples and scaling factors of respectively of 0.5486, 0.1773, 0.8755, 0.6229, 0.9550 times the original signal. Some particular spectra of D&M receiver outputs processing a multipath signal can be found in Figure A-12.

| Label numbers | Delay (# of samples) | | | | | Scaling factor (compared to the original signal) | | | | |
|---------------|-------------------------|-----|-----|-----|-----|---|--------|--------|--------|--------|
| | 1 | 2 | 3 | 4 | 5 | 1 | 2 | 3 | 4 | 5 |
| 11 | 3 | 63 | 139 | 23 | 64 | 0.5486 | 0.1773 | 0.8755 | 0.6229 | 0.9550 |
| 12 | 121 | 114 | 197 | 63 | 186 | 0.9714 | 0.4355 | 0.6787 | 0.2443 | 0.8283 |
| 13 | 95 | 26 | 189 | 144 | 175 | 0.0966 | 0.6231 | 0.1224 | 0.3951 | 1.2890 |
| 14 | 117 | 31 | 171 | 27 | 90 | 0.0026 | 0.6769 | 0.1324 | 0.7804 | 1.0197 |
| 15 | 70 | 29 | 4 | 26 | 47 | 0.3045 | 1.0739 | 0.9110 | 0.8398 | 0.0594 |
| 16 | 144 | 148 | 40 | 74 | 6 | 0.3122 | 0.9214 | 1.2630 | 0.3655 | 0.8747 |
| 17 | 49 | 58 | 143 | 72 | 80 | 0.5342 | 1.0108 | 0.2296 | 1.3243 | 0.5754 |
| 18 | 101 | 20 | 153 | 12 | 20 | 0.9305 | 1.0549 | 1.4575 | 0.9543 | 0.0512 |
| 19 | 36 | 35 | 107 | 194 | 164 | 1.2843 | 0.9261 | 0.2237 | 1.0721 | 0.6657 |

Table 3-6 Multipath labels with their random chosen delay and scaling factor parameters

A significant difference is observable between the output produced by receivers with CC and receivers without CC.

- Receivers not using complex conjugation show spectral lines at the chip rate frequency and frequencies that are multiples of f_c in any of the above defined multipath cases.
- Receivers using complex conjugation that process the above defined multipath signals can show spectra without any spectral lines at the frequencies of interest for the cases where the multipath cases labeled as 12, 16 and 19 are used, or they show only a few spectral lines that are multiples of the chip rate frequency for the multipath cases labeled as 11 and 18. The output spectrum of receivers with CC, and multipath cases labeled as 13 and 14 show some weak spectral lines while for multipath cases labeled as 15 and 17 the effect of multipath is not significant and all spectral lines appear at the output.

3.4 Filtering and Integration

Figure 3-9 shows the block diagram that describes the system used to detect the presence of the flip-wave spread spectrum signal. In Figure A-13, simulation results are shown displaying signals at the output of the different blocks of the system used for signal presence detection. The output of the D&M receiver processing the spread flip-wave signal is followed by a narrow band filter, a squarer, an integrator and a threshold detector. The frequencies for the flip-wave input signal, random data switch, and chip rate are respectively 2000Hz, 400Hz, 256,000Hz or in a 5:1:640 ratio. That means that 5 data bits are contained in 1 data switch interval.

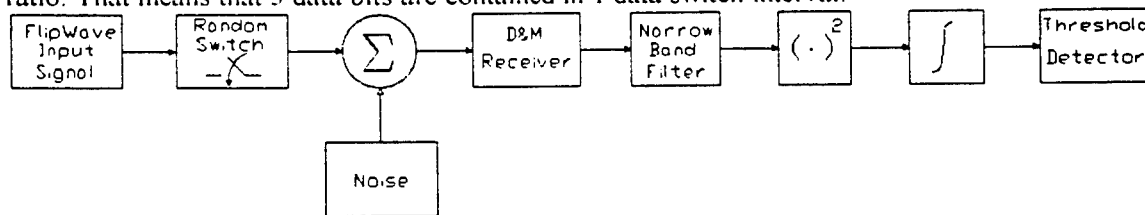


Figure 3-10: Signal detection block diagram system design

Table 3-7 shows the various parameters associated with the signal presence detection problem, such as the required threshold levels, filter bandwidth, noise, and multipath. Several simulations were used to determine effective parameter settings, such as, SNR values of the additive Gaussian noise and the bandwidth of the narrow band filter. The effect of channel induced multipath is also considered. The given threshold levels are obtained after comparing the highs and lows of the peaks of the integrator-detector with the known logical 0's and 1's random switch states for each of the 30 simulated switch state intervals. The separation between the highest 0 and the lowest 1 determines how precisely the threshold level has to be set. A logarithmic separation is defined by

$$\sigma = \text{Log}[\text{Lowest 1}] - \text{Log}[\text{Highest 0}] \quad (3.3)$$

Approximate values are given for the threshold level used to obtain simple relationships of the parameter settings with respect to the needed threshold level.

| SNR (dB) | Band width (Hz) | Multi path (Label) | Thres hold | σ | Logical 0 | | Logical 1 | |
|-------------|--------------------|--------------------------|---------------|----------|-----------|----------|-----------|----------|
| | | | | | lowest | highest | lowest | highest |
| 10 | 100 | n/a | 5e-12 | -0.39 | 7.47e-14 | 2.27e-11 | 9.34e-12 | 2.19e-10 |
| 10 | 1000 | n/a | 2e-10 | 0.15 | 1.34e-12 | 1.75e-10 | 2.45e-10 | 4.55e-09 |
| 10 | 5000 | n/a | 1e-09 | 0.82 | 1.17e-11 | 5.86e-10 | 3.60e-09 | 4.12e-08 |
| 10 | 15000 | n/a | 1e-08 | 1.00 | 5.65e-11 | 2.55e-09 | 2.57e-08 | 2.25e-07 |
| 24 | 100 | n/a | 1e-14 | 2.93 | 1.88e-19 | 1.03e-15 | 8.72e-13 | 6.73e-12 |
| 24 | 1000 | n/a | 1e-12 | 3.68 | 3.37e-18 | 1.33e-14 | 6.37e-11 | 1.93e-10 |
| 24 | 5000 | n/a | 1e-12 | 4.24 | 2.94e-17 | 6.90e-14 | 1.20e-09 | 1.76e-09 |
| 24 | 15000 | n/a | 1e-12 | 4.64 | 1.42e-16 | 2.12e-13 | 9.22e-09 | 1.30e-08 |
| 10 | 100 | 11 | 1e-08 | 0.90 | 8.91e-11 | 1.70e-09 | 1.34e-08 | 1.41e-07 |
| 10 | 1000 | 11 | 1e-07 | 1.41 | 2.47e-09 | 2.18e-08 | 5.63e-07 | 4.08e-06 |
| 10 | 5000 | 11 | 1e-06 | 1.53 | 1.21e-08 | 1.47e-07 | 5.01e-06 | 6.54e-05 |
| 10 | 15000 | 11 | 1e-05 | 1.63 | 5.69e-08 | 6.78e-07 | 2.91e-05 | 6.15e-05 |
| 24 | 100 | 11 | 1e-09 | 1.10 | 2.24e-16 | 4.07e-10 | 5.13e-09 | 4.80e-08 |
| 24 | 1000 | 11 | 1e-08 | 1.67 | 6.12e-15 | 5.12e-09 | 2.39e-07 | 9.16e-07 |
| 24 | 5000 | 11 | 1e-07 | 1.98 | 3.04e-14 | 2.53e-08 | 2.39e-06 | 1.63e-05 |
| 24 | 15000 | 11 | 1e-06 | 2.21 | 1.43e-13 | 7.67e-08 | 1.23e-05 | 1.35e-04 |
| 24 | 100 | 12 | 5e-10 | 1.73 | 6.03e-17 | 1.50e-10 | 7.97e-09 | 6.67e-08 |
| 24 | 1000 | 12 | 1e-08 | 2.14 | 2.95e-15 | 2.11e-09 | 2.90e-07 | 9.38e-07 |
| 24 | 5000 | 12 | 1e-07 | 2.02 | 2.75e-14 | 1.98e-08 | 2.08e-06 | 7.75e-06 |
| 24 | 15000 | 12 | 1e-06 | 1.81 | 1.38e-13 | 1.54e-07 | 9.99e-06 | 5.57e-05 |

Table 3-7 Threshold level extracted from logical 0's and 1's

The range of integrated output values for logical 0's is bigger (on a logarithmic scale) than that for the logical 1's. The use of complex conjugation in the D&M Receiver does not setting the threshold level. Only the spectrum (of the D&M output) appears different when referring to the spectral line at 0 Hz. Two different multipath cases labeled as number 11 and number 12 are included in the threshold level simulations. These two multipath cases have an effect on the performance of the D&M receiver that is different in each case. In analyzing the performance of signal presence detection by threshold detection, the different values of the two multipath system parameters do not seem to affect the performance. The values set for the threshold level to detect signal presence are shown in the graph of Figure 3-11. Figure 3-12 shows the separation defined by Equation 3.3 for simulated cases.

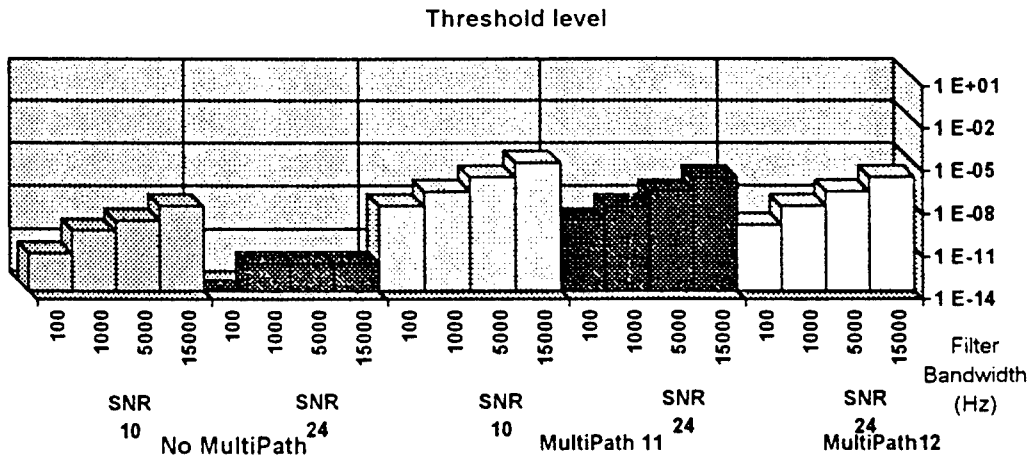


Figure 3-11: Used Threshold level for the analyzed systems

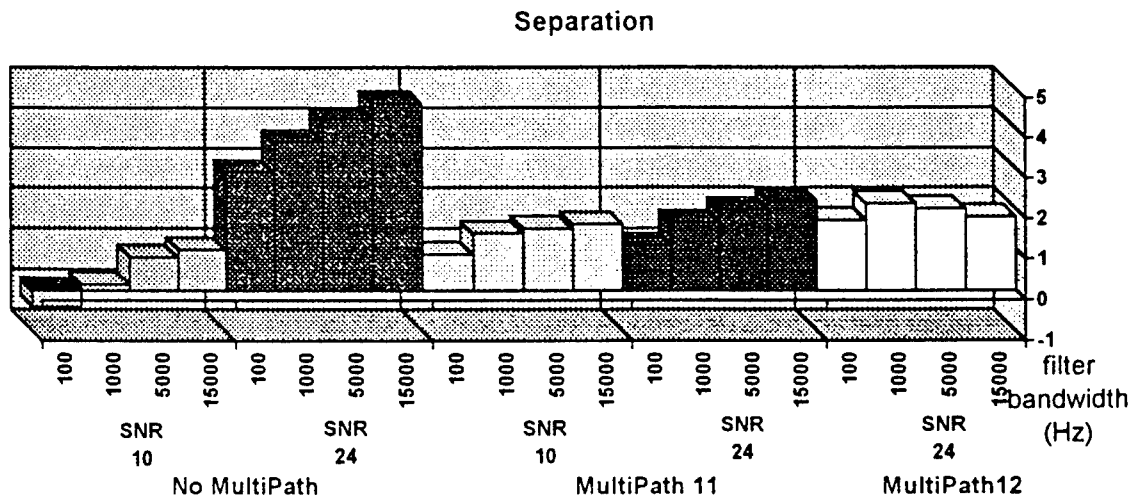


Figure 3-12: Separation available for selecting threshold level

A negative separation indicates that the highest logical 0 observed after the integration is higher than the lowest observed logical 1. In that case the threshold detector can not be set to a level that results in a 100% error free signal presence detection. Using a larger filter bandwidth and higher SNR, in general enlarges the separation and reduces the likelihood of signal presence detection errors.

4 Conclusions and Recommendations

The features of a delay and multiply receiver are described in Section 1.3. (More details are found in [1, 5]). The performance of delay and multiply receivers processing low probability of intercept signals (described in Section 2) is tested under different conditions in order to study:

1. the effect of the delay setting of the D&M receiver varying from 0 to 1 chip duration interval. This is evaluated for signals spread by various shapes of HBE pulses imposed on PN sequence generator AM signals.
2. the effect of noise added to the input signal that is processed by the D&M receiver.
3. the effect of multipath on the performance of the D&M receiver processing an input signal that is degraded by channel induced multipath.

Furthermore, post processing of the D&M receiver output by a combination of a narrow band filter, squarer, and integrator that is used for signal presence detection is investigated from a performance standpoint.

The delay of the receiver is evaluated for DSSS signals with raised cosine, rectangular, and several sinc shaped HBE pulses. The variation in the spectral line strength at the chip rate frequency for the different shaped HBE pulses confirm that an appropriate delay setting in the D&M receiver depends partly on the kind of HBE pulse used.

The use of complex conjugation affects the performance of the receiver definitely under the various operational scenarios studied.

Not only the magnitude of the spectral lines are significant when it comes to signal detection, but also the shape of the total spectrum, and in particular the power and bandwidth of the main lobe are affected. White Gaussian noise added to the system makes the simulations more realistic, but does not significantly affect the performance of the receiver as long as specific signal to noise ratios are maintained. Interference in the form of multipath propagation makes the performance of the delay and multiply receiver unpredictable. However the long observation time detector in the form of the threshold detector using post D&M receiver processing is a remarkably reliable signal presence detector if the threshold level is appropriately set. The performance of the threshold detector depends strongly on the filter bandwidth used and noise level of the system. The presence of multipath has sometimes a stabilizing effect on the signal presence detection performance. Therefore, further studies into automatic threshold level settings for given bandwidth and SNR values are necessary to completely specify the implementation of such a post D&M receiver processor.

References

- [1] Bukofzer, D. C., *Performance Analysis and Simulation Results of Delay and Multiply Receivers Processing a Spread Spectrum Modulated Flip-wave-Signal Generated from High Bandwidth Efficiency Pulses*, Final Report for Summer Faculty Research Program, Rome Laboratory, August 1996.
- [2] Dixon, Robert, *Spread Spectrum Systems*, 2nd edition, John Wiley & Sons, New York, N.Y., 1984, ISBN-0-471-88309-3
- [3] Gill, W.H., and Spilker, J.J., *An Interesting Decomposition Property for the Self-Products of Random or Pseudorandom Binary Sequences*, IEEE Trans. Comm., Vol. COM-11, June 1963, pp. 246-247
- [4] Kamillo Feher, DR., *Advanced Digital Communications, Systems and Signal Processing Techniques*, 2nd edition, Prentice-Hall, New Jersey, 1987, ISBN-0-13-011198-8
- [5] Kuehls, J.F., and Geraniotis, E., *Presence Detection of Binary-Phase-Shift-Keyed and Direct-Sequence Spread-Spectrum Signals Using a Prefilter Delay-and-Multiply Device*, IEEE Journ. On Selected Areas in Comm., Vol. 8, No. 5, June 1990, pp. 915-933.

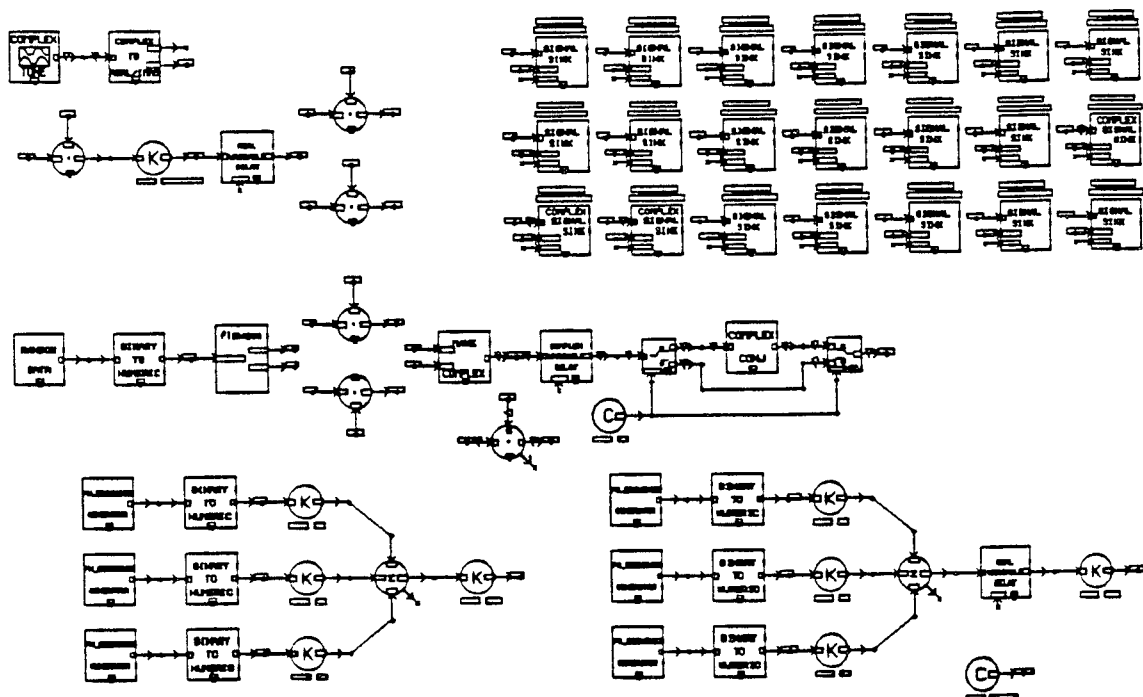


Figure A-1: Block Diagram Design (BDD): Transmitter and Receiver system

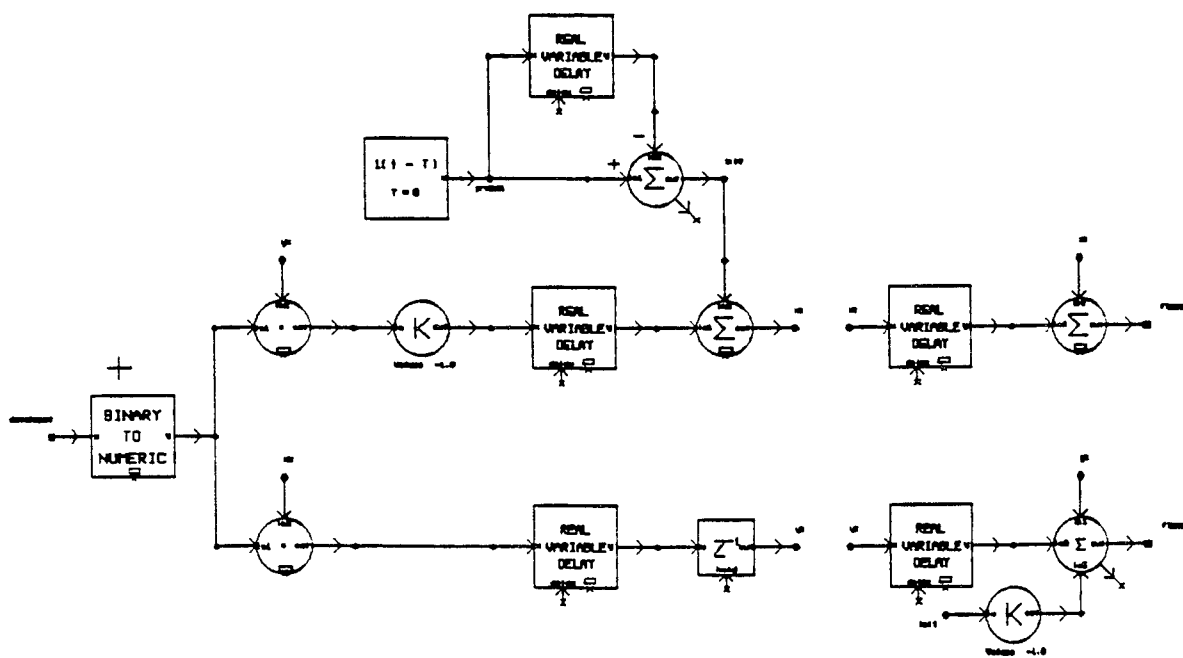


Figure A-2: Block Diagram Design (BDD): The flip-wave generator

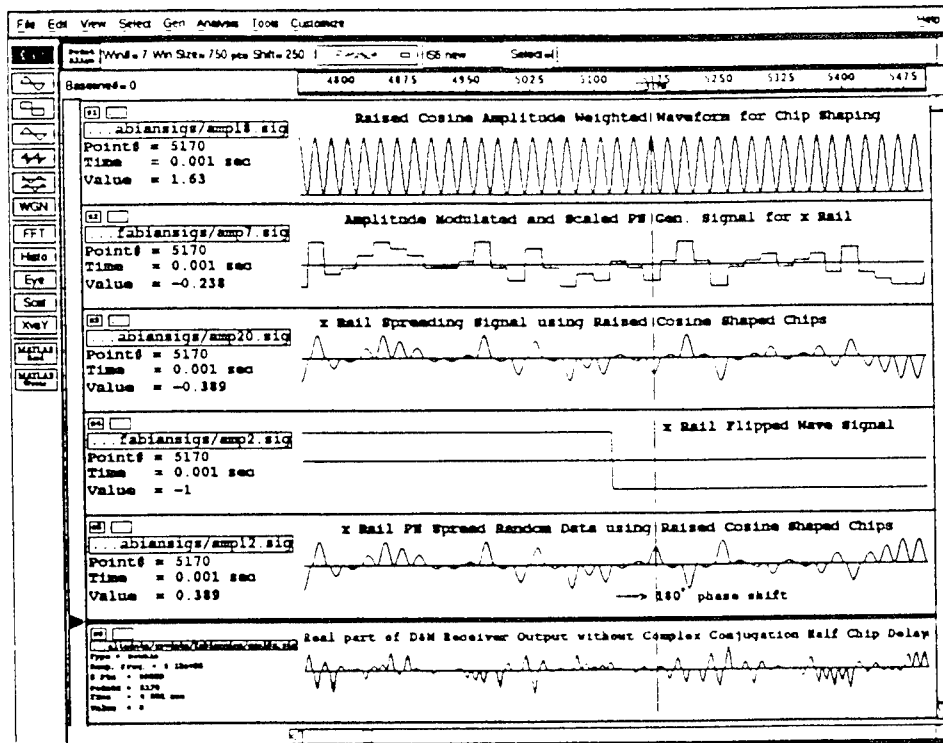


Figure A-3: Signal Simulation: PN spreading of random data with raised cosine

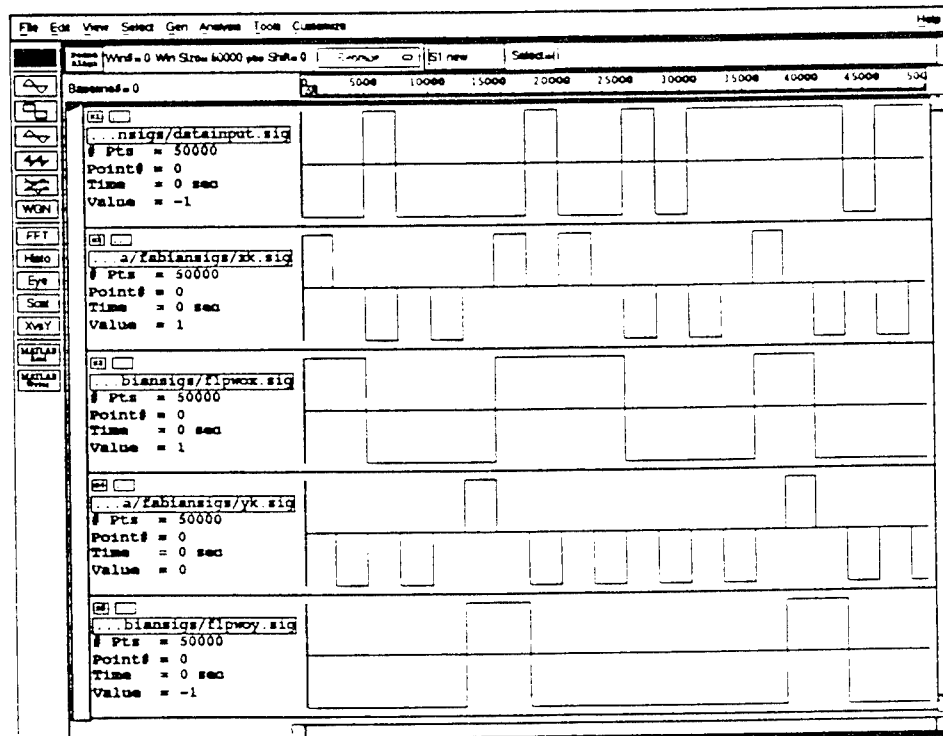
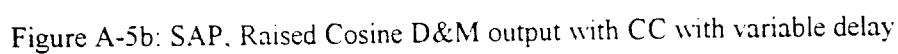
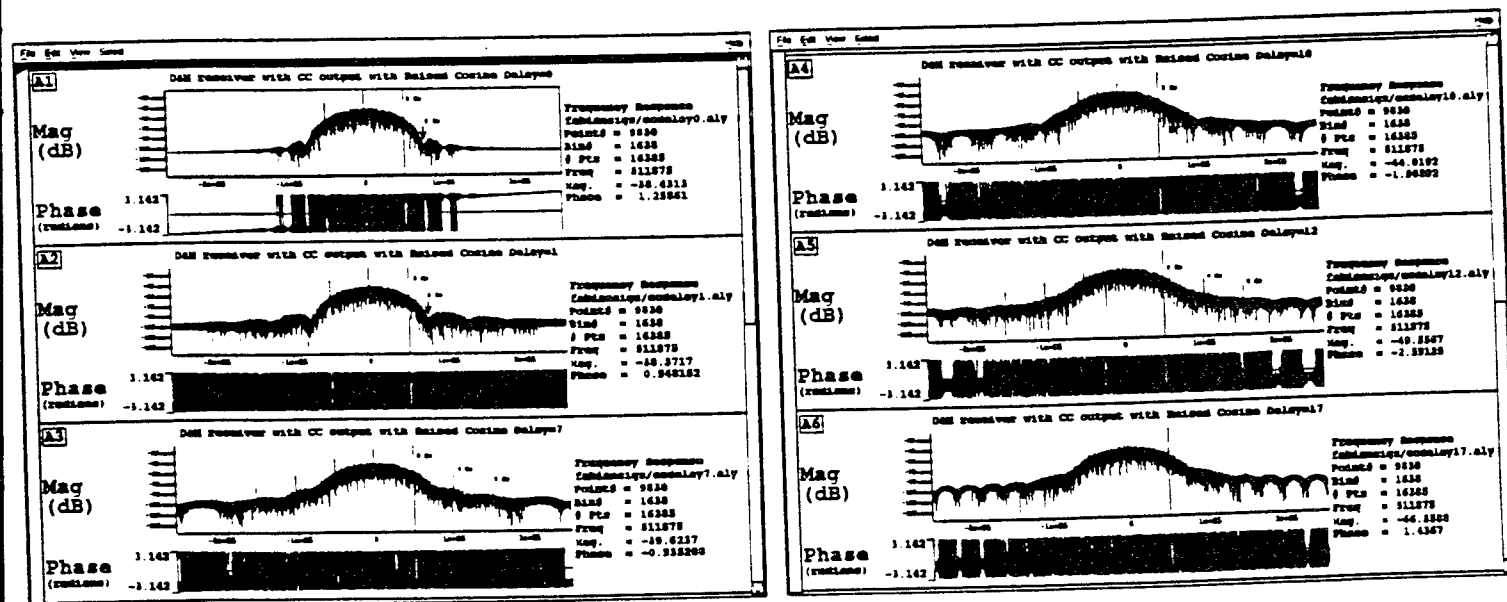
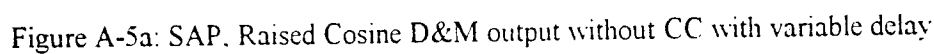
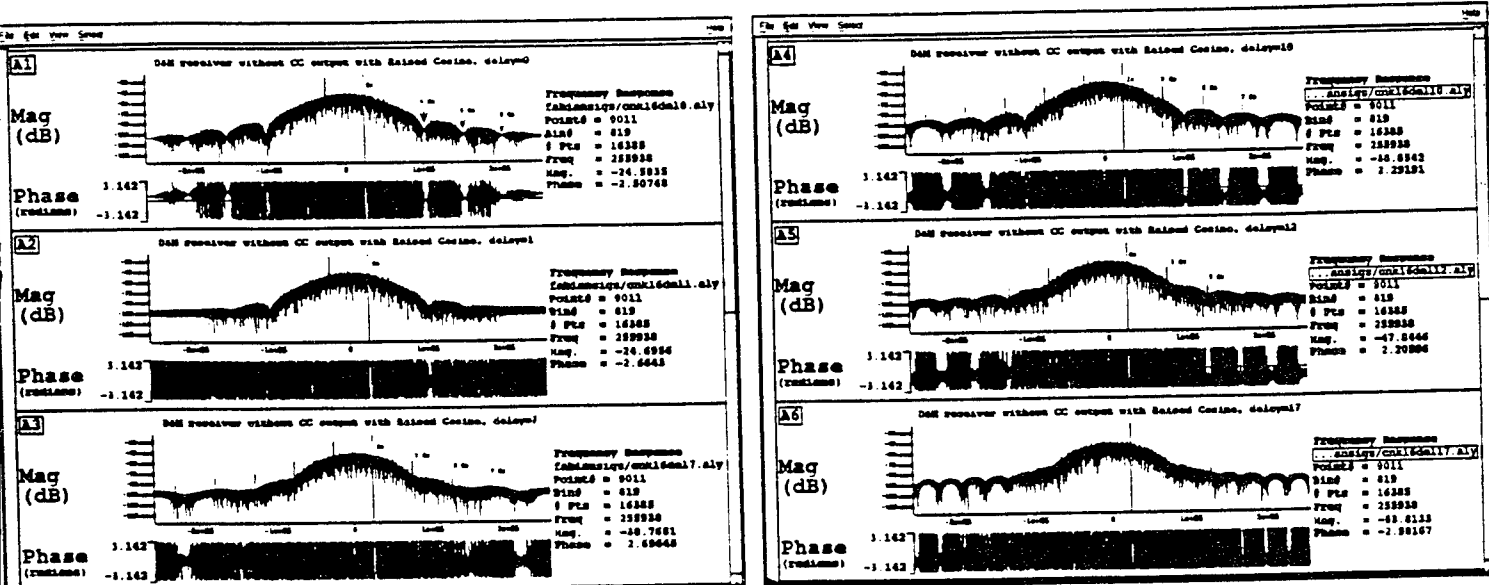


Figure A-4: Signal Simulation: Flip-wave signal



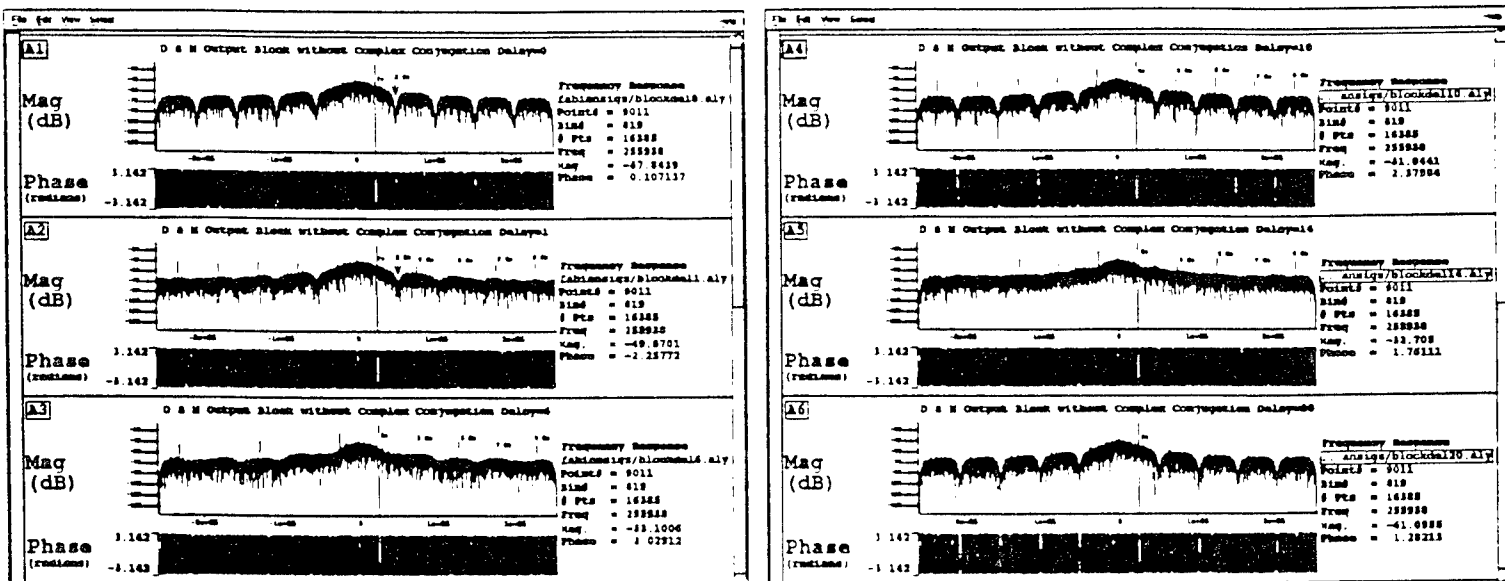


Figure A-6a: SAP. Rectangular shaped pulses D&M output without CC with variable delay

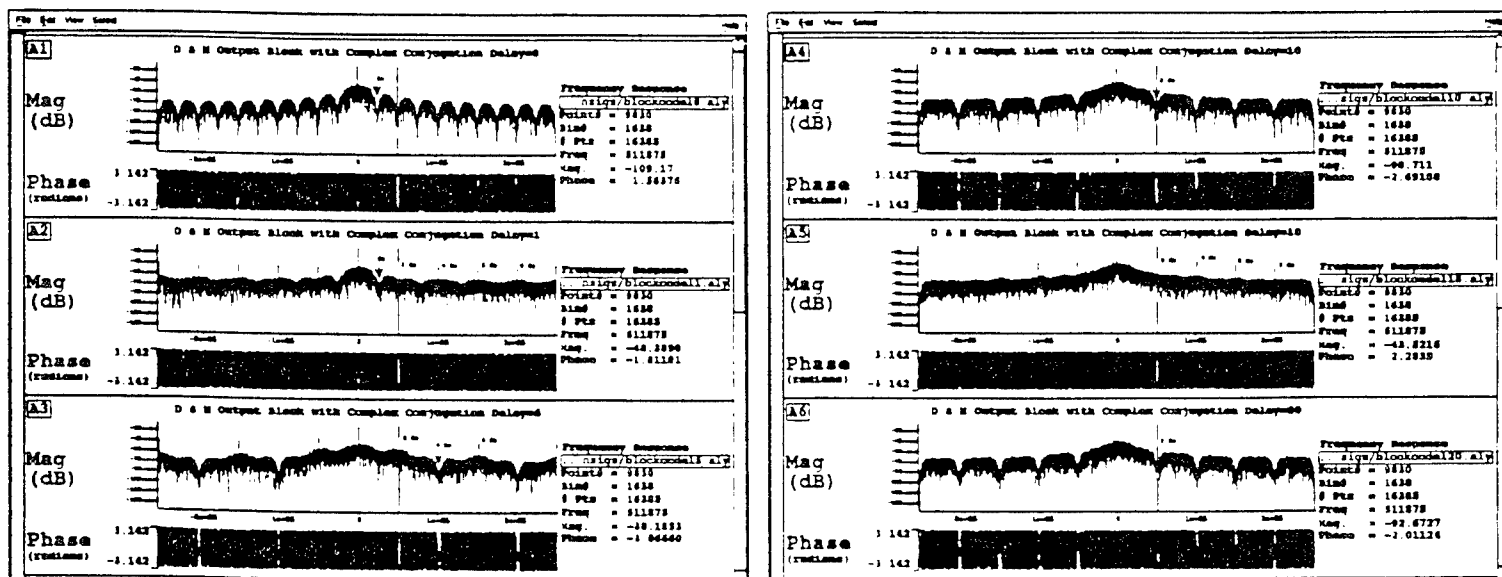


Figure A-6b: SAP. Rectangular shaped pulses D&M output with CC with variable delay

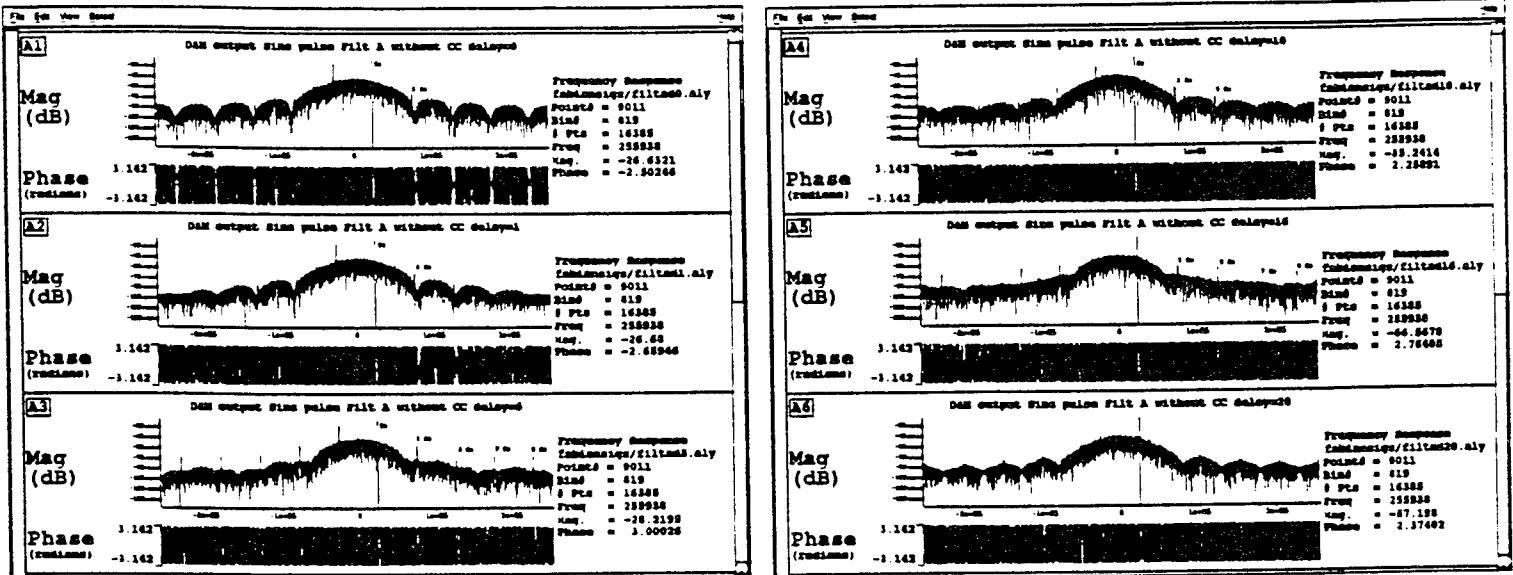


Figure A-7a: SAP, Sinc shaped pulses of duration $T=4T_c$ D&M output with variable delay

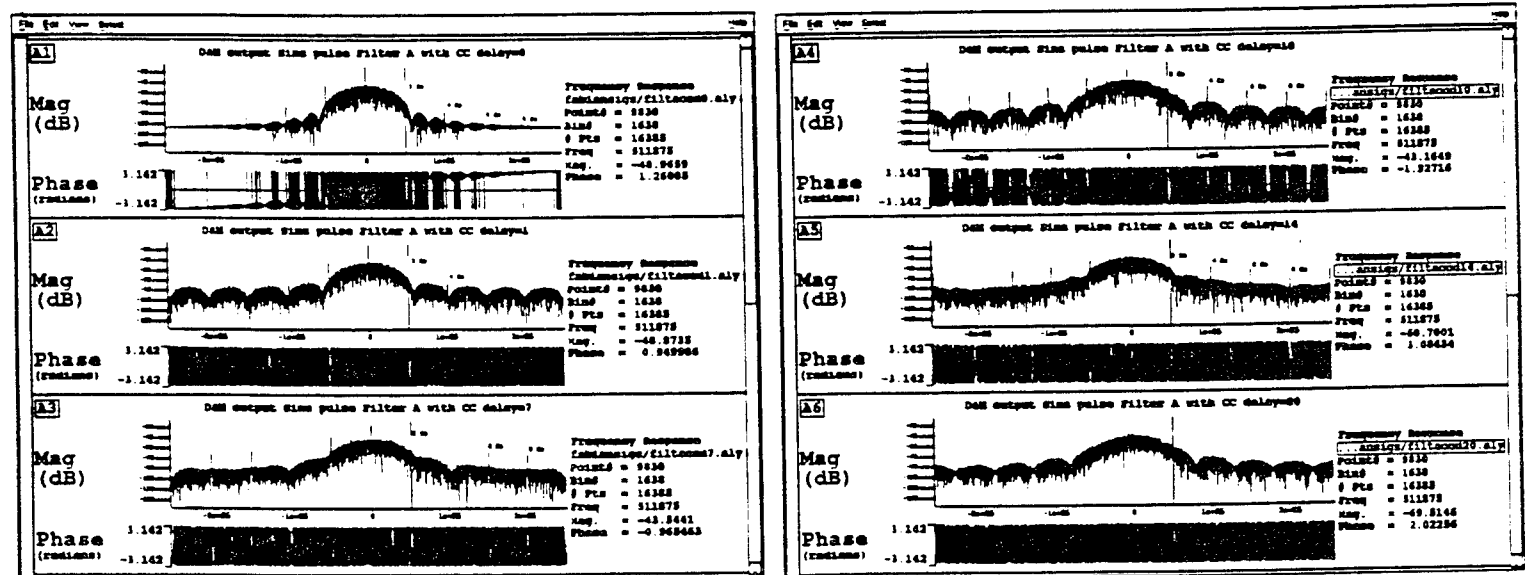


Figure A-7b: SAP, Sinc shaped pulses of duration $T=4T_c$ D&M output with variable delay

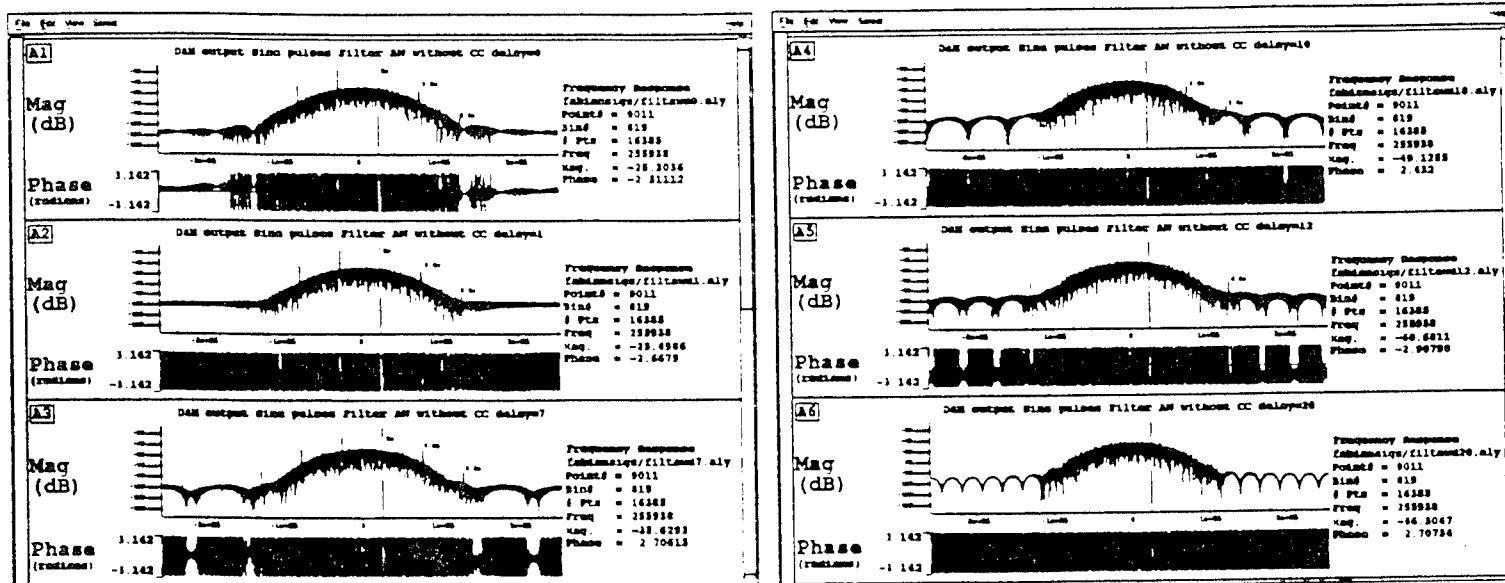


Figure A-7c: SAP, Sinc shaped pulses of duration $T=4T_c$ D&M output with variable delay

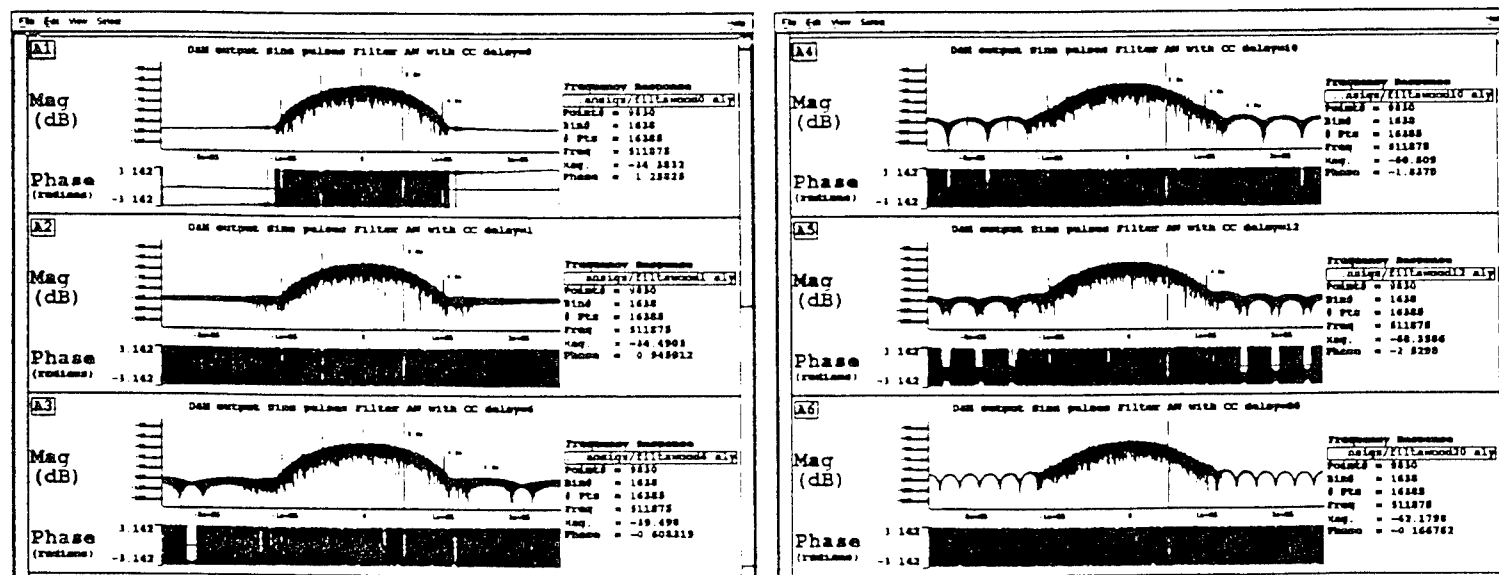


Figure A-7d: SAP, Sinc shaped pulses of duration $T=4T_c$ D&M output with variable delay

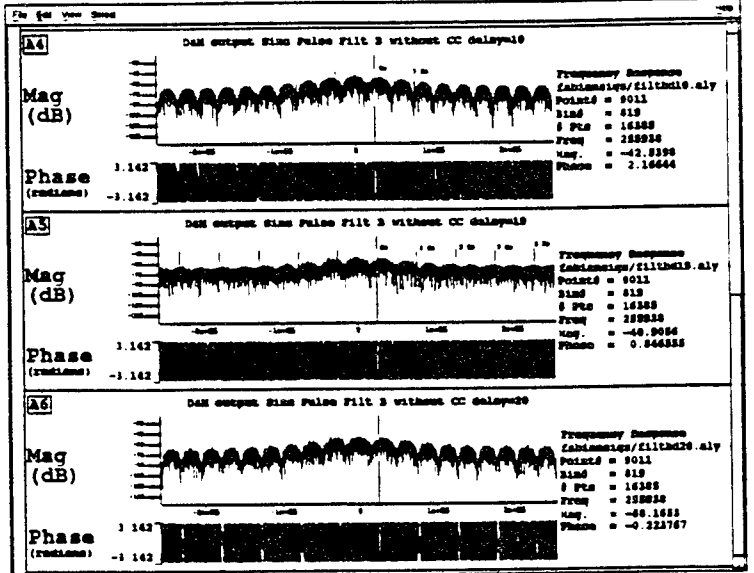
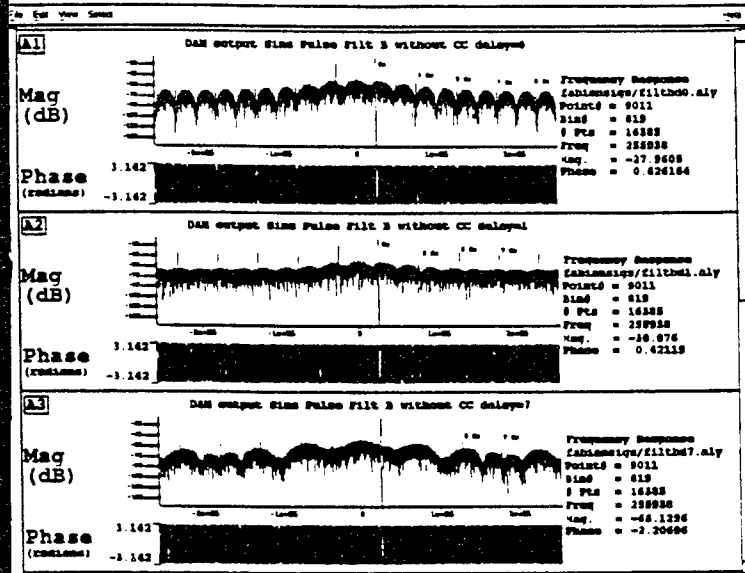


Figure A-8a: SAP, Sinc shaped pulses of duration $T=8T_c$ D&M output with variable delay

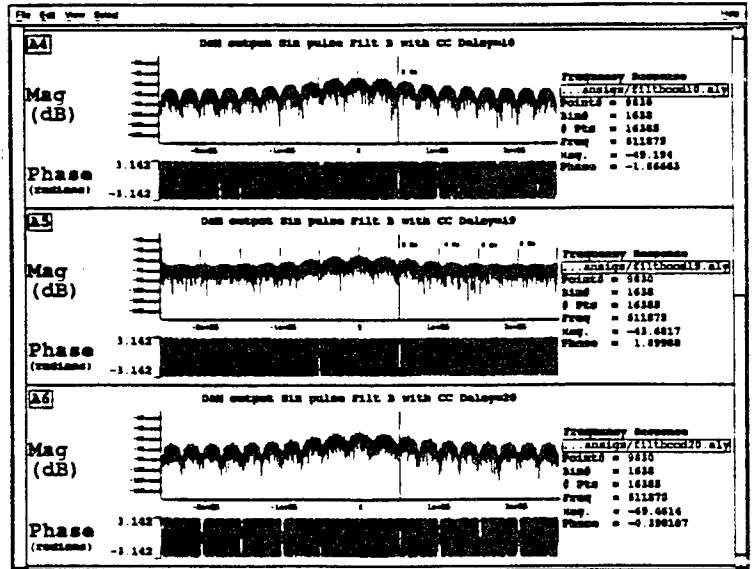
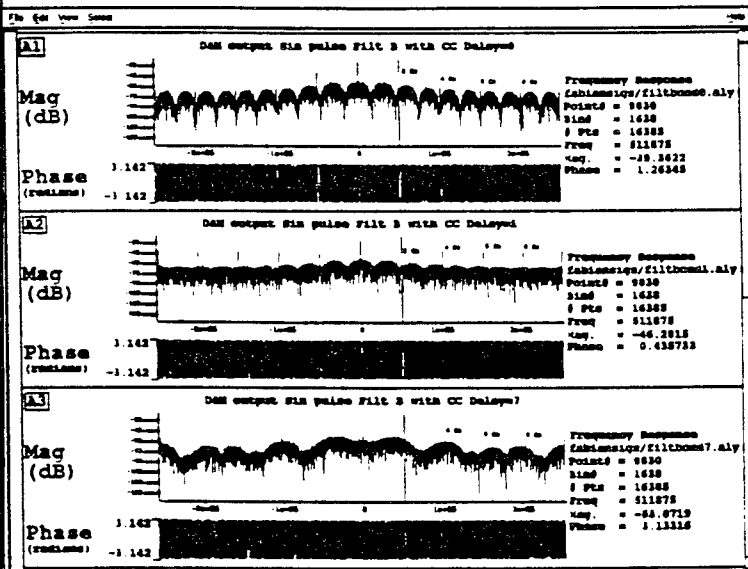


Figure A-8b: SAP, Sinc shaped pulses of duration $T=8T_c$ D&M output with variable delay

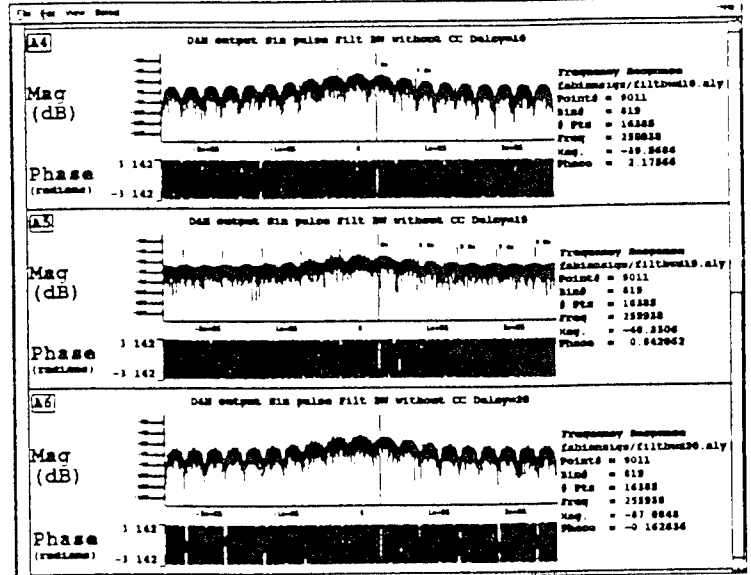
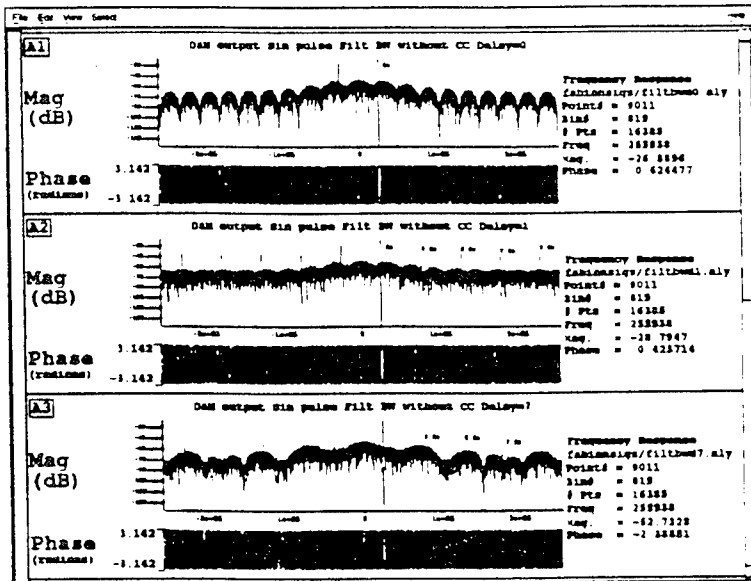


Figure A-8c: SAP, Sine shaped pulses of duration $T=8T_c$ D&M output with variable delay

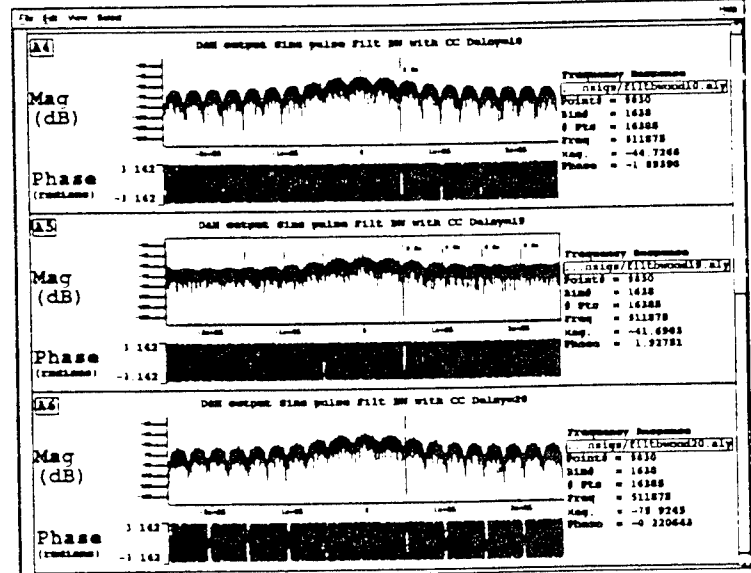
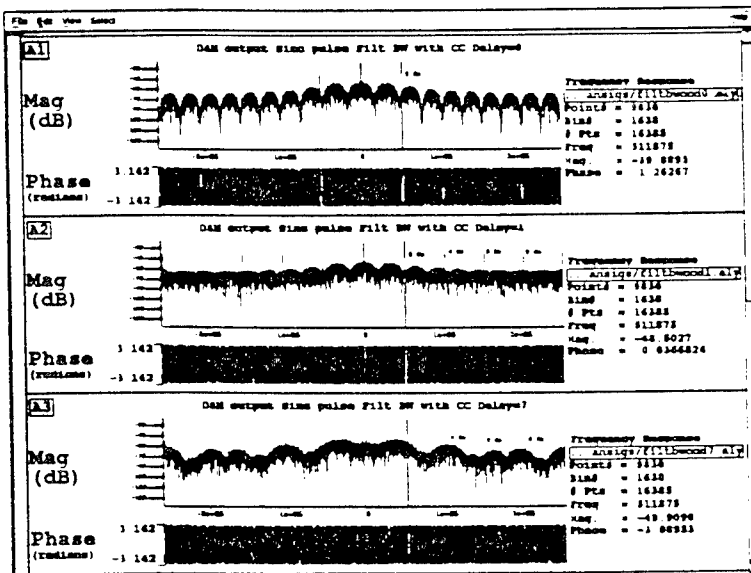


Figure A-8d: SAP, Sine shaped pulses of duration $T=8T_c$ D&M output with variable delay

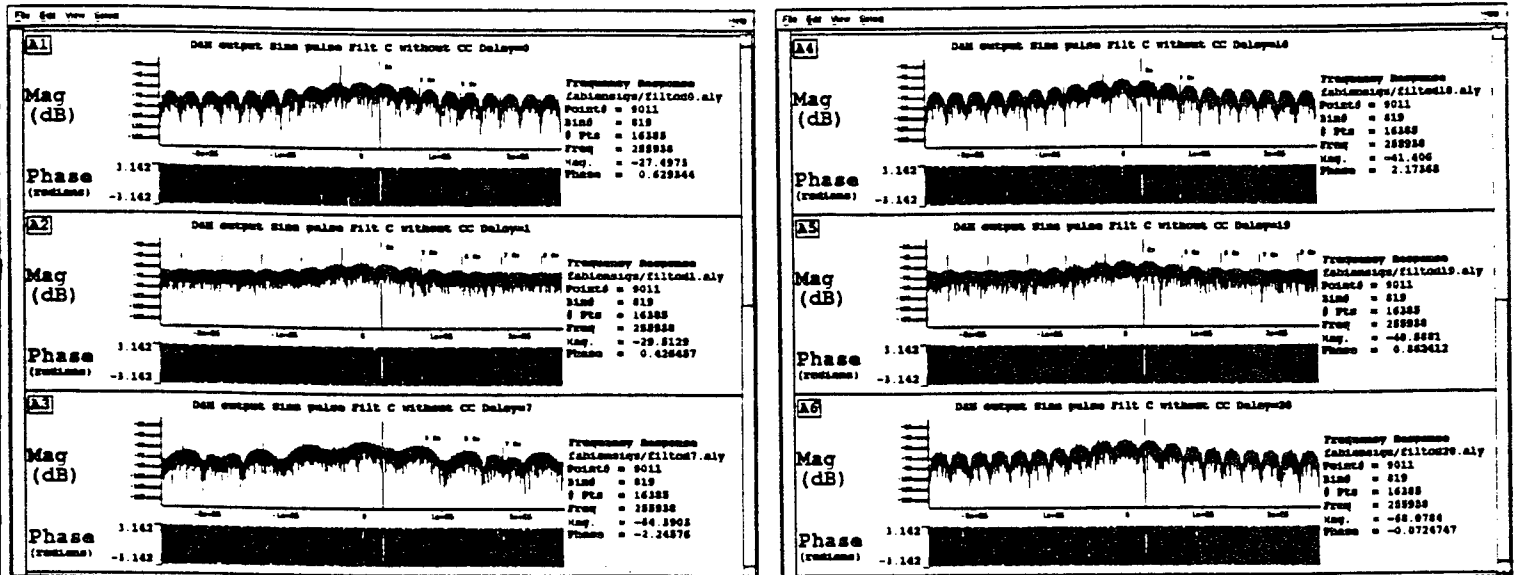


Figure A-9a: SAP, Sinc shaped pulses of duration $T=16T_c$ D&M output with variable delay

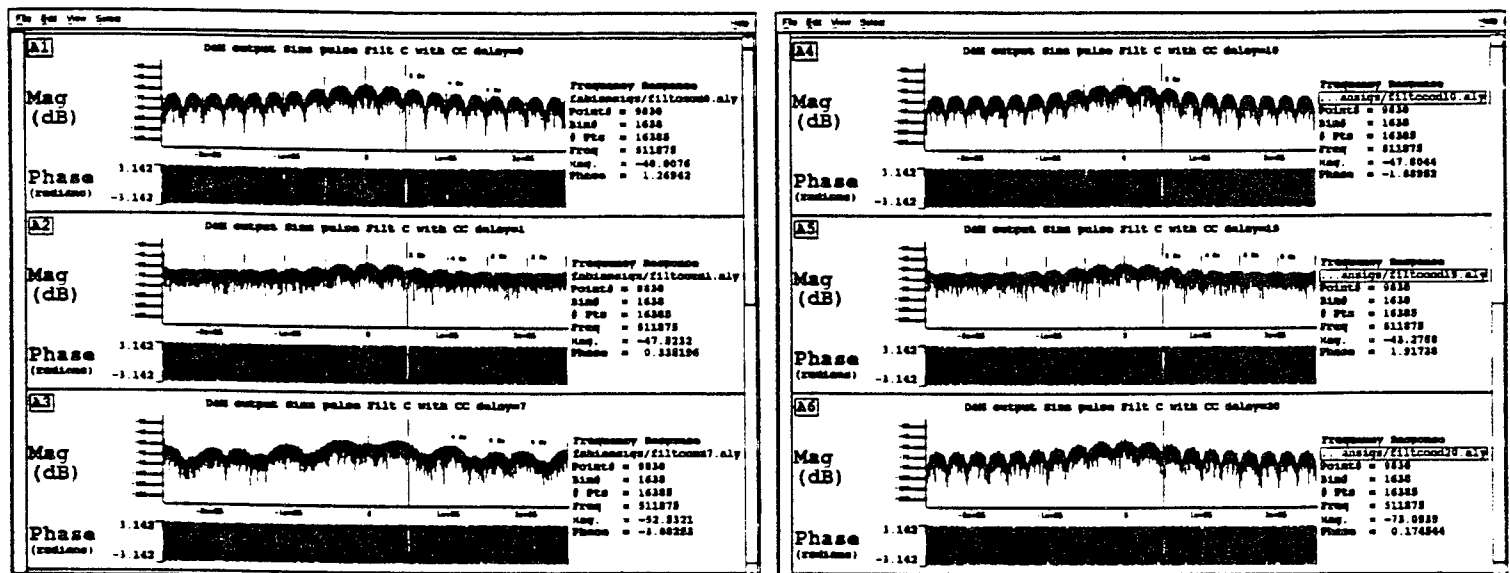


Figure A-9b: SAP, Sinc shaped pulses of duration $T=16T_c$ D&M output with variable delay

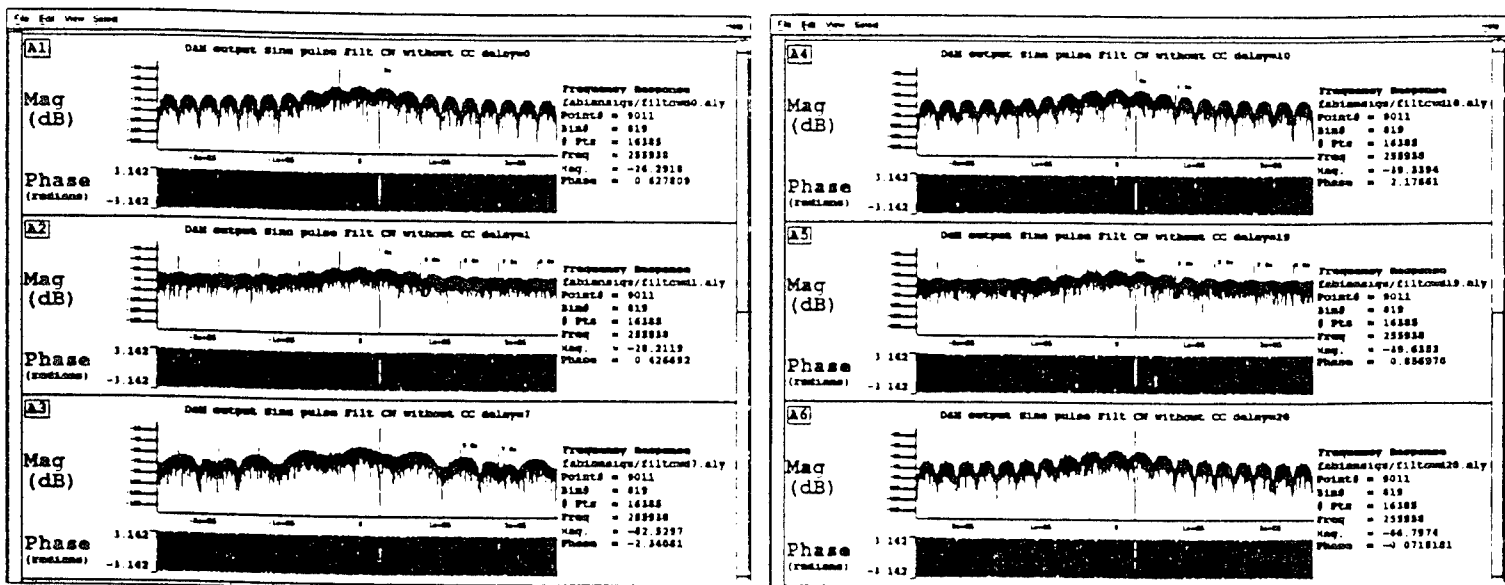


Figure A-9c: SAP, Sinc shaped pulses of duration $T=16T_c$ D&M output with variable delay

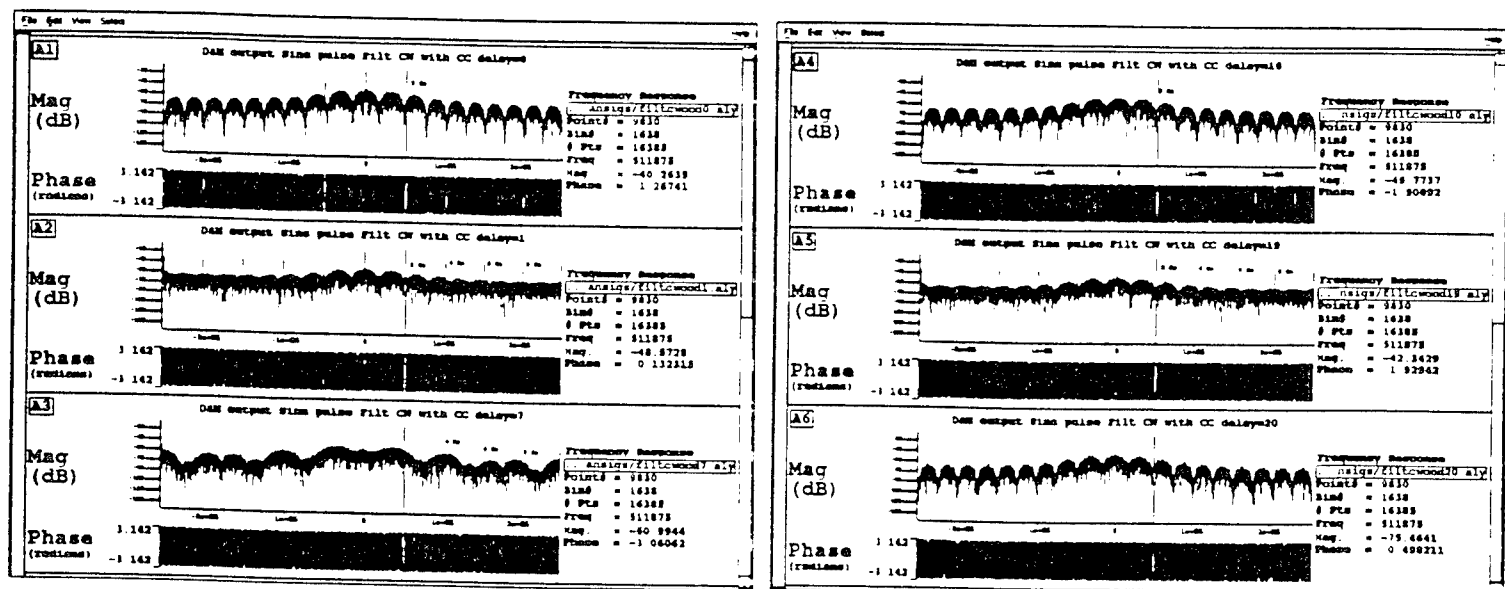


Figure A-9d: SAP, Sinc shaped pulses of duration $T=16T_c$ D&M output with variable delay

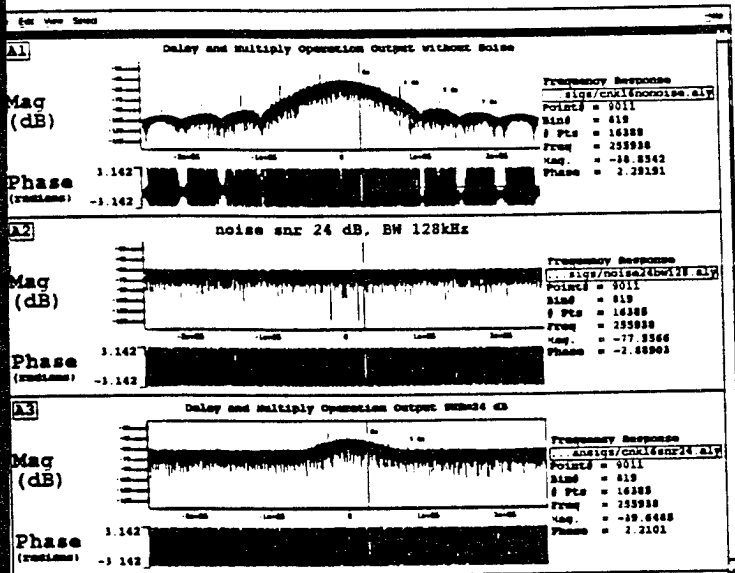


Figure A-10: SAP, Noise simulation

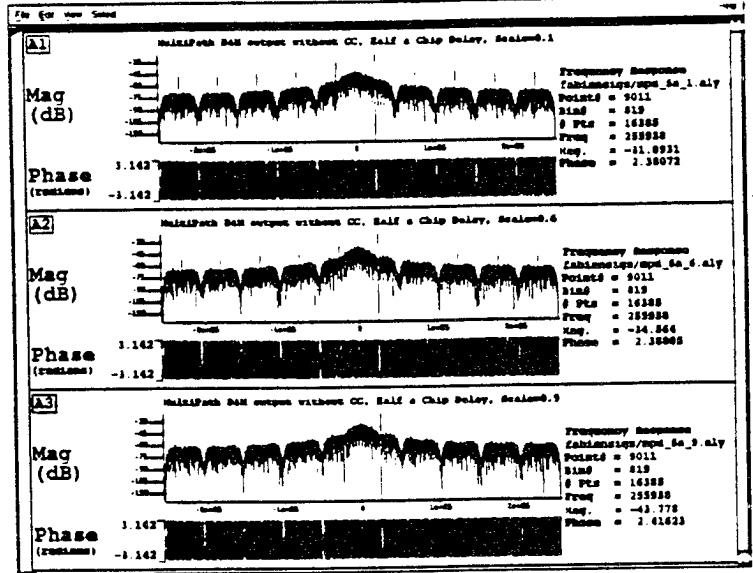


Figure A-11: SAP, Multipath with one delayed and scaled duplicate

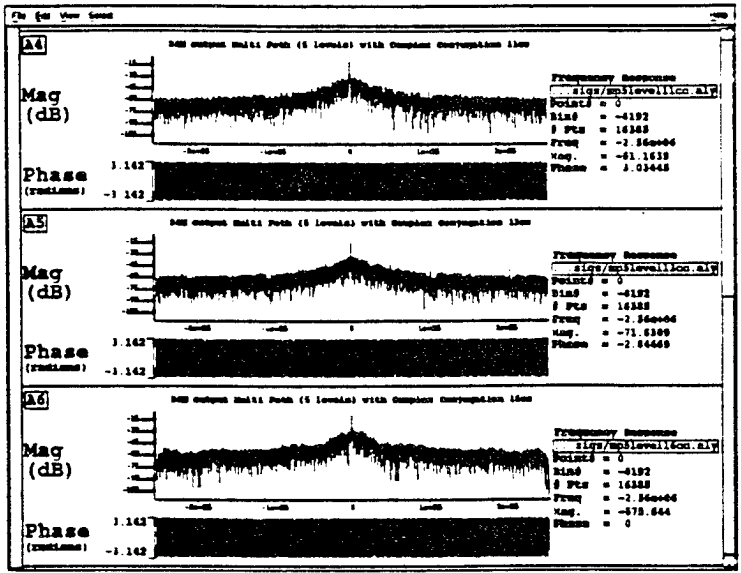
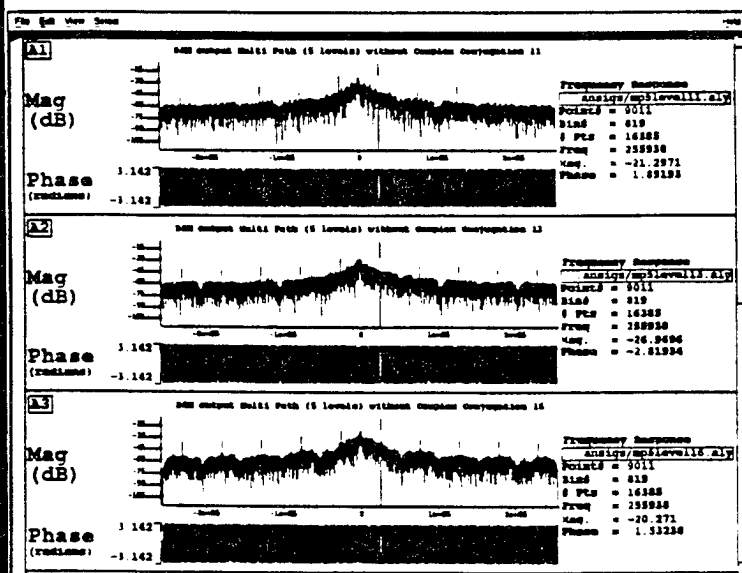


Figure A-12: SAP, Multipath with 5 delayed and scaled duplicates

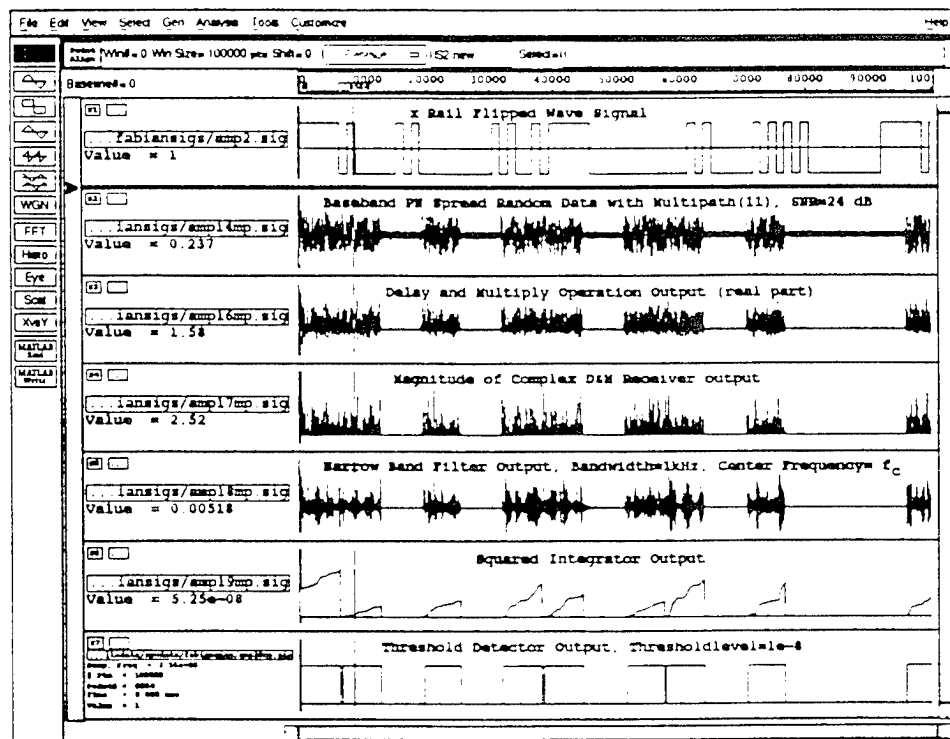


Figure A-13: Signal Simulation: Threshold detection

**NON-DESTRUCTIVE AND OPTICAL CHARACTERIZATION OF COMPOSITION
AND THICKNESS IN MULTILAYER TERNARY SEMICONDUCTOR STACKS**

Dr. Xuesheng Chen, Assistant Professor
Department of Physics and Astronomy
Wheaton College
East Main St., Norton, MA 02766
Office: (508)-286-3977
E-mail address: xchen@wheatonma.edu

Final Report for
Summer Research Extension Program (SREP)
(United States AFOSR Contract #: F49620-93-C-0063,
Subcontract#: 97-0887)

Sponsored by
Air Force Office of Scientific Research
Bolling Air Force Base, DC
and
Wheaton College

March 1998

NON-DESTRUCTIVE AND OPTICAL CHARACTERIZATION OF COMPOSITION AND THICKNESS IN MULTILAYER TERNARY SEMICONDUCTOR STACKS

Xuesheng Chen, Assistant Professor
Department of Physics and Astronomy
Wheaton College
East Main St., Norton, MA 02766

Abstract

Multilayer $\text{In}_{1-x}\text{Ga}_x\text{As}/\text{InP}$ structures have found wide application in high-speed electronic and optical devices. The composition x , the layer thickness, and their uniformity in the structure are crucial in obtaining desirable device performance. In this report, we first present the spectorelectance technique to determine the composition x and the layer thickness of $\text{In}_x\text{Ga}_{1-x}\text{As}$ in a multilayer stack. Then, we describe the photoluminescence method to determine the composition and its uniformity across the wafer, and to describe the important role photoluminescence can play in obtaining an accurate refractive index function $n(x,\lambda)$ for $\text{In}_{1-x}\text{Ga}_x\text{As}$.

NON-DESTRUCTIVE AND OPTICAL CHARACTERIZATION OF COMPOSITION AND THICKNESS IN MULTILAYER TERNARY SEMICONDUCTOR STACKS

Xuesheng Chen

Introduction

Most semiconductor devices are optimized by heterojunctions, which are commonly achieved through the use of ternary semiconductor epitaxial layers such as $\text{In}_x\text{Ga}_{1-x}\text{As}$. Epitaxial layers are the building blocks in optoelectronic device fabrication. Due to its superior electronic properties, the ternary semiconductor $\text{In}_x\text{Ga}_{1-x}\text{As}$ has found wide applications in high-speed electronic and optical devices such as p-i-n detectors, avalanche photodiodes, and long wavelength diode lasers. The ternary composition x is crucial in obtaining desirable device performance, and the uniformity of each epitaxial layer affects device yields. Routine ternary composition and thickness measurements by TEM, optical microscopy and double crystal X-ray diffraction (DCXRD) are usually destructive, labor intensive, and time-consuming when applied to evaluate large area uniformity. The optical reflectance mapping method developed by Weyburne and his collaborators at Air Force Research Laboratory - Hanscom AFB has recently been shown to have this kind of desirable property for AlAs/GaAs and $\text{Al}_{1-x}\text{Ga}_x\text{As/GaAs}$ multilayer systems [1,2]. I joined Dr. Weyburne's group during the summer of 1996, as an AFOSR Summer Faculty Research Associate, and extended this method to $\text{In}_{1-x}\text{Ga}_x\text{As/InP}$ multilayer stacks to see if the composition, thickness and their uniformity of each layer can be determined accurately. It is essential with this method to have a model to describe the dependence of the refractive index n on the composition x and photon wavelength λ for $\text{In}_x\text{Ga}_{1-x}\text{As}$ because the optical reflectance depends on $n(x,\lambda)$ and layer thickness in the stack. In literature, there are models for $n(x,\lambda)$, but only for lattice matched composition $x=0.53$. During the summer 1996, I developed a model for $n(x,\lambda)$ by modifying Jensen's model [3]. Comparing with modified Adachi's model [4], however, I found that the deduced composition and thickness for a

$\text{In}_x\text{Ga}_{1-x}\text{As}$ layer in the stack from optical reflectance are model dependent. There are not enough reliable experimental indices available in the literature in our composition and wavelength range for us to determine which model is correct. Here, in this work, we (I was collaborating with Dr. Weyburne and Dr. Paduano) show that it is possible to determine the refractive indices n or effective n for the epitaxial layer $\text{In}_x\text{Ga}_{1-x}\text{As}$ on InP substrate in the composition x and wavelength λ range we are interested in, and to obtain a reliable formula to describe $n(x,\lambda)$, using combination of spectroreflectance, photoluminescence and X-ray diffraction. In this report, we first describe the optical reflectance technique to determine the composition and thickness of $\text{In}_x\text{Ga}_{1-x}\text{As}$ in the multilayer stack. Then, we describe the photoluminescence (PL) method to determine the composition and its uniformity across the wafer, and the important role photoluminescence can play in determine accurate $n(x,\lambda)$.

Spectroreflectance

A. Theoretical Reflectance

The reflectance R from a multilayer stack such as $\text{In}_x\text{Ga}_{1-x}\text{As} / \text{InP}$ is given by the equation below [5]:

$$R = \left| \frac{\eta_o - \frac{Y}{E}}{\eta_o + \frac{Y}{E}} \right|^2, \quad (1)$$

$$\begin{bmatrix} E \\ Y \end{bmatrix} = \left(\prod_{j=1}^n M_j \right) M_c \left(\prod_{j=1}^m M_j \right) \begin{bmatrix} I \\ \eta_s \end{bmatrix}, \quad (2)$$

where the matrices M_j are given by:

$$M_j = \begin{bmatrix} (\exp i\delta_j + \exp -i\delta_j)/2 & (\exp i\delta_j - \exp -i\delta_j)/2\eta_j \\ \eta_j(\exp i\delta_j - \exp -i\delta_j)/2 & (\exp i\delta_j + \exp -i\delta_j)/2 \end{bmatrix} \quad (3)$$

The meanings of those symbols in eqs.(1) to (3) are the following:

j denotes the j-th layer,

s denotes the substrate,

o denotes the air,

c denotes the middle cavity layer in a multilayer stack,

n denotes number of pairs (e.g. $\text{In}_x\text{Ga}_{1-x}\text{As} / \text{InP}$) above the cavity,

m denotes number of pairs below the cavity and above the substrate,

$\delta_j = 2\pi N_j d_j \cos \theta_j / \lambda$, (λ is wavelength, and d_j is the j-th layer thickness, θ_j is the incident angle to interface of the j-th and (j+1)-th layers),

$N_j = n - ik$, (n and k are the j-th layer refractive index and extinction coefficients, respectively),

$\eta_j = N_j \cos \theta_j$ for TE mode or $\eta_j = N_j / \cos \theta_j$ for TM mode.

As you can see, the reflectance R depends on refractive index $n(x, \lambda)$, thickness d of each layer, and the wavelength λ . If $n(x, \lambda)$ is known for each epitaxial layer in the multilayer stack, its composition x and thickness d can be deduced by treating them as adjustable parameters to get the best fit of the experimental reflectance curve to the theoretical reflectance R that was described by eq.(1).

Fig.1 shows the simulated reflectance curve with three different compositions by using the modified Adachi's model [2,4] for $n(x, \lambda)$. The curve shifts obviously to the up and the right with small increment in composition and is very sensitive to the composition change for a stack similar to the structure shown in Fig.2. It was shown that the optical reflectance is also highly sensitive to the thickness of each epitaxial layer in a multilayer AlAs/GaAs stack [1]. The simulated reflectance curves for the stack with three different GaAs layer thickness are

shown in Fig.3. The dashed lines indicate the effects of increasing the GaAs layer thickness by +0.5, +1.0, and +2.0%, respectively.

B. Results and Discussions

The work was done on multilayer $\text{In}_x\text{Ga}_{1-x}\text{As}/\text{InP}$ stacks that were grown by MOCVD at AFRL, Hanscom AFB. Undoped InP and $\text{In}_x\text{Ga}_{1-x}\text{As}$ layers were grown using trimethylaluminum (TMI), triethylgallium (TEG), phosphine, and arsine as sources. Single side polished, (100)-oriented, semi insulating InP substrates were used. The test structures consist of 7 or 9 pairs of $\text{In}_x\text{Ga}_{1-x}\text{As}/\text{InP}$ and a half wavelength cavity layer of $\text{In}_x\text{Ga}_{1-x}\text{As}$ (see Fig.2). The typical thickness for the pair is 130 nm and 160 nm for $\text{In}_x\text{Ga}_{1-x}\text{As}$ and InP, respectively, and the cavity has a thickness of 260 nm. To avoid light adsorption by the layers, we choose the wavelength range of 1600 nm to 2200 nm which is mostly in the transparent region just above the $\text{In}_x\text{Ga}_{1-x}\text{As}$ bandgap wavelength. The measurement range is kept near the $\text{In}_x\text{Ga}_{1-x}\text{As}$ bandgap so that the index of reflection will have significant wavelength dependence. This insures that the fitted composition will be unique. Choosing a center wavelength around 1900 nm means that the quarter-wave InP stack thickness and the $\text{In}_x\text{Ga}_{1-x}\text{As}$ stack thickness can be calculated by

$$n d = \lambda_{\text{center}} / 4 \quad , \quad (4)$$

where d is the layer thickness, λ_{center} is the center wavelength, and n is the appropriate index of refraction evaluated at λ_{center} . The half-wave InGaAs cavity layer on top of the stack is grown to be twice the InGaAs stack thickness.

The index of refraction of lattice-matched InGaAs was calculated according to Adachi [4]. The parameters were adjusted to account for small variation in the composition around this value by fitting the experiment data to Adachi's model. The index of InP is obtained by fitting data from Palik [7] (1000-2200 nm) to a semi-empirical Sellmerier type equation as shown below:

$$n_{\text{InP}} = \left[6.948 + \frac{2.585}{1 - \frac{382903.75}{\lambda^2}} \right]^2, \quad (5)$$

where λ is in nm. Then the experimental reflectance curve was fitted to eq.(1) to deduce the composition and thickness. The experimental and fitted reflectance spectra for 7-pair and 9-pair $\text{In}_x\text{Ga}_{1-x}\text{As}/\text{InP}$ structures are shown in Figs.4 and 5. The minimum at the center of the high reflectance zone corresponds to the cavity resonance. This minimum location is a good approximation of λ_{center} and provides a convenient method of estimating initial values for the spectrum fitting. The deduced composition x and layer thickness are shown in Figs. 6 and 7 for a 7-pair stack. In terms of composition variation with respect to the lattice-matched composition, a variation range of -0.2% to +0.7% is observed from the center to the edge of the wafer. For 85% of the total area at the center of wafer, the thickness variation is < 0.7%. Excluding 4 mm from the edge of the wafer, the uniformity is within $\pm 0.5\%$. Therefore, this $\text{In}_x\text{Ga}_{1-x}\text{As}$ layer is lattice matched to InP substrate over 90% of the wafer area.

Spectroreflectance method described here is very sensitive to the composition and thickness as long as a structure similar to Fig.2 is used. The accuracy of the thickness, and especially the composition (x), however, depends on the model for $n(x,\lambda)$. See Appendix for details.

Photoluminescence Method

It was shown that the wavelength position at the half of photoluminescence (PL) peak intensity shifts with the composition x for $\text{In}_{1-x}\text{Ga}_x\text{As}$ [7]. It would be easy and non-destructive if we can use this wavelength position to determine the composition x and its uniformity of the $\text{In}_{1-x}\text{Ga}_x\text{As}/\text{InP}$ stack. In order to find out how exactly the half-PL-peak intensity wavelength position changes with composition x , PL spectra from more than a dozen of single-layer $\text{In}_{1-x}\text{Ga}_x\text{As}$ layer on InP substrate were measured. The composition x for each of these samples was determined by double crystal X-ray diffraction (see Fig.8). Fig.9 shows PL spectra from

3 samples with different x ($x_1=54.5\%$, dash; $x_2=52.2\%$, solid; $x_3=49.4\%$, dots). The notch in the center of the PL spectra is due to a notch in the grating. Samples are excited by a diode laser at 674nm with 10mW. As you can see, the wavelength position at half PL-peak intensity on the lower energy side changes a lot with composition x . A curve describing the dependence of this wavelength position on composition x is obtained from the PL data, shown in Fig.10.

Double crystal X-ray diffraction (DCXD) can not easily determine composition x in a $\text{In}_{1-x}\text{Ga}_x\text{As}/\text{InP}$ stack. It is, however, very convenient to measure PL and use Fig.10 to determine the x . To see how accurate with this method, PL was measured for a 9-pair $\text{In}_{1-x}\text{Ga}_x\text{As}/\text{InP}$ stack and x was then determined from Fig.10 to be 52.6%, comparing with 52.8% determined by DCXD. It is hard to determine which number is more accurate because DCXD cannot provide composition directly for a multilayer stack (assumptions of each layer thickness have to be made). The composition uniformity map for a 9-pair $\text{In}_{1-x}\text{Ga}_x\text{As}/\text{InP}$ stack using PL method is shown in Fig.11. The PL peak intensity map is shown in Fig.12.

It seems that the composition can be determined easily and accurately using PL method if Fig.10 is accurate. Accuracy of Fig.10 depends on if the compositions of single-layered $\text{In}_{1-x}\text{Ga}_x\text{As}$ on InP substrate can be accurately determined by X-ray. Also PL and X-ray have to be done on the same spot of the sample. PL should be obtained by exciting the sample with a low power light source because the wavelength position at half-PL-peak intensity changes with the excitation power that is above certain limit. Fig.13 shows that PL shapes, positions, and width hardly change when excited by a diode laser with powers below 16.5mW. Fig.14, however, shows that they do change when excited by a Argon laser with powers that range from 4 to 250mW. To be accurate, same type of low-power excitation should be consistently used for the establish of Fig.10 and for PL measurements on to-be-examined stacks.

The PL method does not seem to be able to determine easily the layer thickness in the stack. The spectrophlectance is, however, very sensitive to the thickness, as shown in Fig.3. The spectrophlectance technique relies on an accurate formula of $n(x,\lambda)$, which can be obtained from reflectivity and PL. PL can be used to determine composition x of $\text{In}_{1-x}\text{Ga}_x\text{As}$,

and the reflectance can be used to determine a formula of $n(\lambda)$. Using samples with different x , we should be able to develop an accurate refractive index function $n(x, \lambda)$.

Acknowledgement

Work is supported by AFOSR SREP grant (Contract is F49620-93-C-0063, subcontract is 97-0887). Work is also partially supported by Wheaton College. I am very grateful to Dr. Weyburne and Dr. Paduano at AFRL for their fully support in this work.

Appendix

It is essential to have a theoretical model to describe the dependence of the refractive index n on the composition x and the wavelength λ for $\text{In}_x\text{Ga}_{1-x}\text{As}$ in order to use eq.(1). We developed a model for $n(x, \lambda)$ of $\text{In}_x\text{Ga}_{1-x}\text{As}$ by modifying B. Jensen's model [3]. It uses a quantum mechanics calculations of the dielectric constant of a compound semiconductor and assumes the band structure of Kane Theory. The theoretical expressions for $n(x, \lambda)$ is given in terms of the basic material parameters of band gap energy E_g , effective electron mass m_n , effective hole mass m_p , spin orbit splitting energy Δ , and lattice constant a . In the nonabsorbing range, the $n(x, \lambda)$ of $\text{In}_x\text{Ga}_{1-x}\text{As}$ can be described by

$$n^2 = 1 + 2C_0 \{ (Y_B - Y_F) - z(\tan^{-1}(Y_B / z) - \tan^{-1}(Y_F / z)) \} \quad , \quad (6)$$

where

$$Y_B = m_0(a - a_0) \quad ,$$

$$m_0 = 2.93 \text{ A} \quad ,$$

$$a = (1-x) a_{\text{GaAs}} + x a_{\text{InAs}} \quad ,$$

$$a_{\text{GaAs}} = 5.6534 \text{ A}, \quad a_{\text{InAs}} = 6.0585 \text{ A} \quad ,$$

$$z = [1 - (\hbar\omega/E_g)]^{1/2} \quad ,$$

$$\omega = 2 \pi (c/\lambda), \quad (c = \text{speed of light})$$

$$E_g = 1.43 - 1.53x + 0.45 x^2,$$

$$C_0 = (\omega_v^2/\omega_g^2),$$

$$\omega_g = E_g/\hbar,$$

$$\omega_v = 4\pi e^2 N_v^*/m_n,$$

$$N_v^* = N_v(m_r/m_n)^{3/2},$$

$$N_v = 8/3\pi^2 \chi_c^3,$$

$$\chi_c = \hbar/(E_g m_n/2)^{1/2},$$

$$1/m_r = 1/m_n + 1/m_p,$$

$$m_n = 0.07(1-x)m_e + 0.028xm_e,$$

$$m_p = 0.5(1-x)m_p + 0.33xm_p,$$

$$m_e = 9.10939 \times 10^{-28} \text{ gram},$$

$$Y_F = 2(n_e/N_v)^{1/3},$$

$$n_e = 6.5 \times 10^{16} \text{ cm}^{-3}, \quad (\text{carrier concentration}).$$

All the formulas above for $n(x,\lambda)$ are in cgs units.

Plugging eq.(6) into eq.(1), the theoretical reflectance $R(x,\lambda)$ for a multilayer $\text{In}_x\text{Ga}_{1-x}\text{As}/\text{InP}$ stack can be obtained.

The composition x and the thickness of the $\text{In}_x\text{Ga}_{1-x}\text{As}$ and InP in the 7 pairs and the thickness of the $\text{In}_x\text{Ga}_{1-x}\text{As}$ cavity can be easily found by treating them as adjustable parameters to get the best fit of the experimental reflectivity curve to the theoretical curve described by eq.(1). For a sample of a 7-pair $\text{In}_x\text{Ga}_{1-x}\text{As}(\sim 130\text{nm})/\text{InP}(\sim 160\text{nm})$ with a half $\text{In}_x\text{Ga}_{1-x}\text{As}$ cavity on top, we obtained from the fitting:

$$\text{thickness } d_{\text{InGaAs in pairs}} = 1378\text{\AA}, \quad \text{thickness } d_{\text{InP in pairs}} = 1547\text{\AA},$$

thickness $d_{\text{InGaAs cavity}}=2553\text{\AA}$, composition $x_{\text{InGa}_{1-x}\text{As}} = 0.555$.

These values are close to the targeted growth values.

However, when the same experimental curve was fitted to the Adachi's model [6], we obtained:

thickness $d_{\text{InGaAs in pairs}}=1414\text{\AA}$, thickness $d_{\text{InP in pairs}}=1502\text{\AA}$,
thickness $d_{\text{InGaAs cavity}}=2557\text{\AA}$, composition $x_{\text{InGa}_{1-x}\text{As}} = 0.578$.

The two sets of the deduced values are different, especially in the composition, due to the two different models for the $n(x,\lambda)$ of $\text{In}_x\text{Ga}_{1-x}\text{As}$.

References

- [1] Qing S. Paduano, David Weyburne, Fenglu and R. Bhat, J. Elect. Mat. Vol.24, No.11,1995, p1659.
- [2] Qing S. Paduano and David Weyburne, J. Appl. Phys. 1998.
- [3] B. Jensen and A. Torabi, J. Appl. Phys. Lett. Vol.33, No.7, 1978, p659.
- [4] S. Adachi, J. Appl. Phys. Vol.53, No.8, 1982, p5863.
- [5] H. A. MaLeod, Thin-Film Optical Filters (Bristol: Hilger, 1986), p32.
- [6] E. Palik, Handbook of Optical Constants of Solids, (Academic Press, Boston, 1985) p. 512.
- [7] I. C. Bassignane, C. J. Miner, and N. Puetz, J. Appl. Phys. Vol.65, No.11, 1 June 1989, p4299.

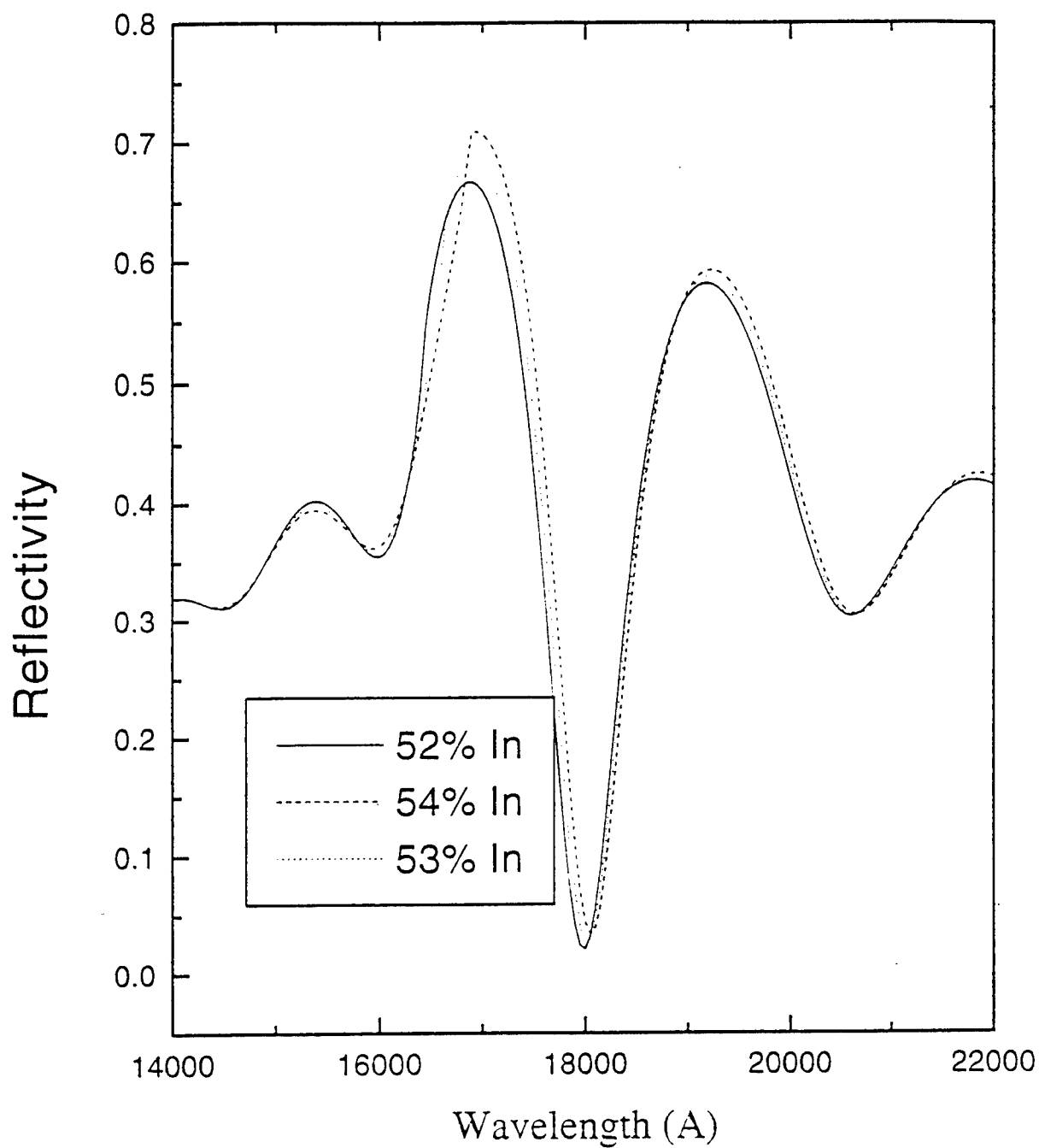


Fig.1 Simulated reflectivity curves for a $\text{In}_x\text{Ga}_{1-x}\text{As}/\text{InP}$ multilayer stack with three different In compositions using the modified Adachi's model [2,4] for the refractive index $n(x,\lambda)$.

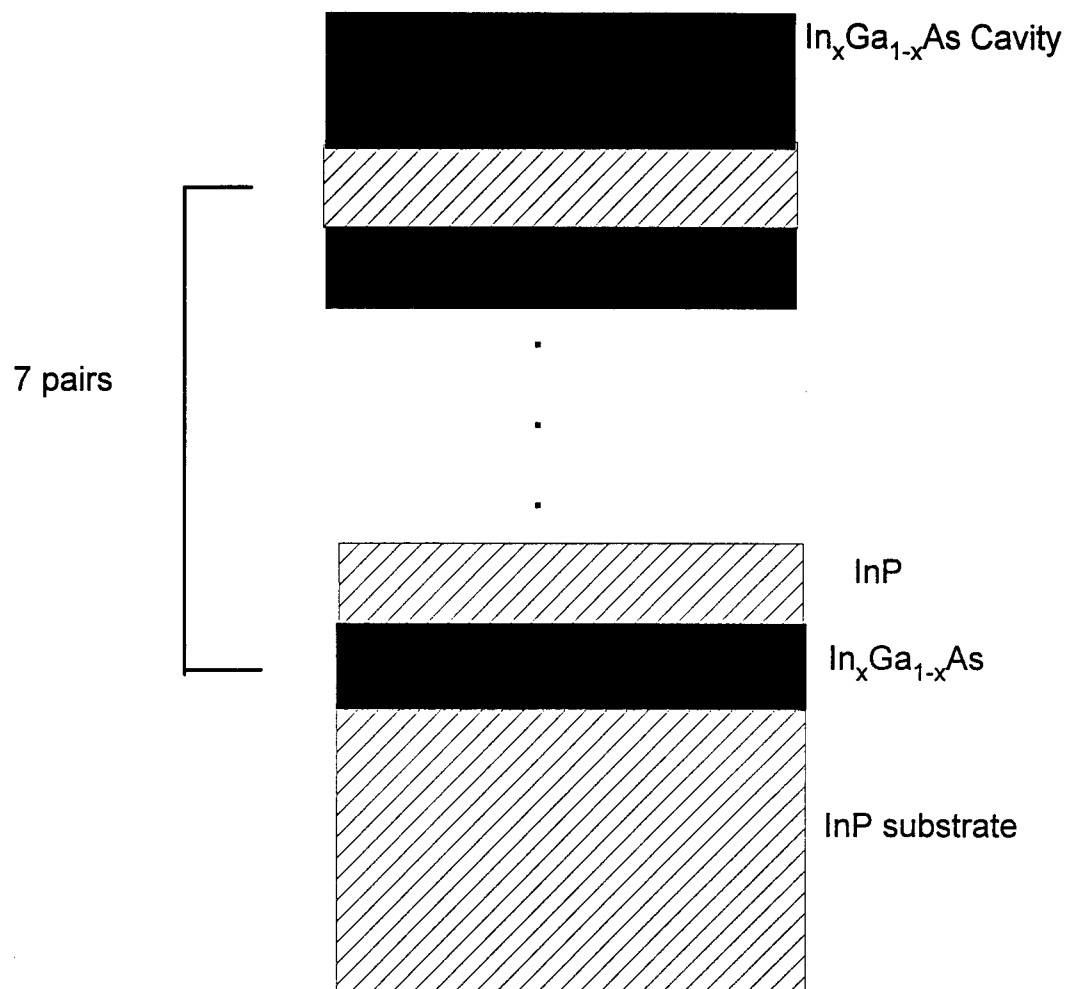


Fig.2. The cross section of the $\text{In}_x\text{Ga}_{1-x}\text{As}/\text{InP}$ multilayer stack.

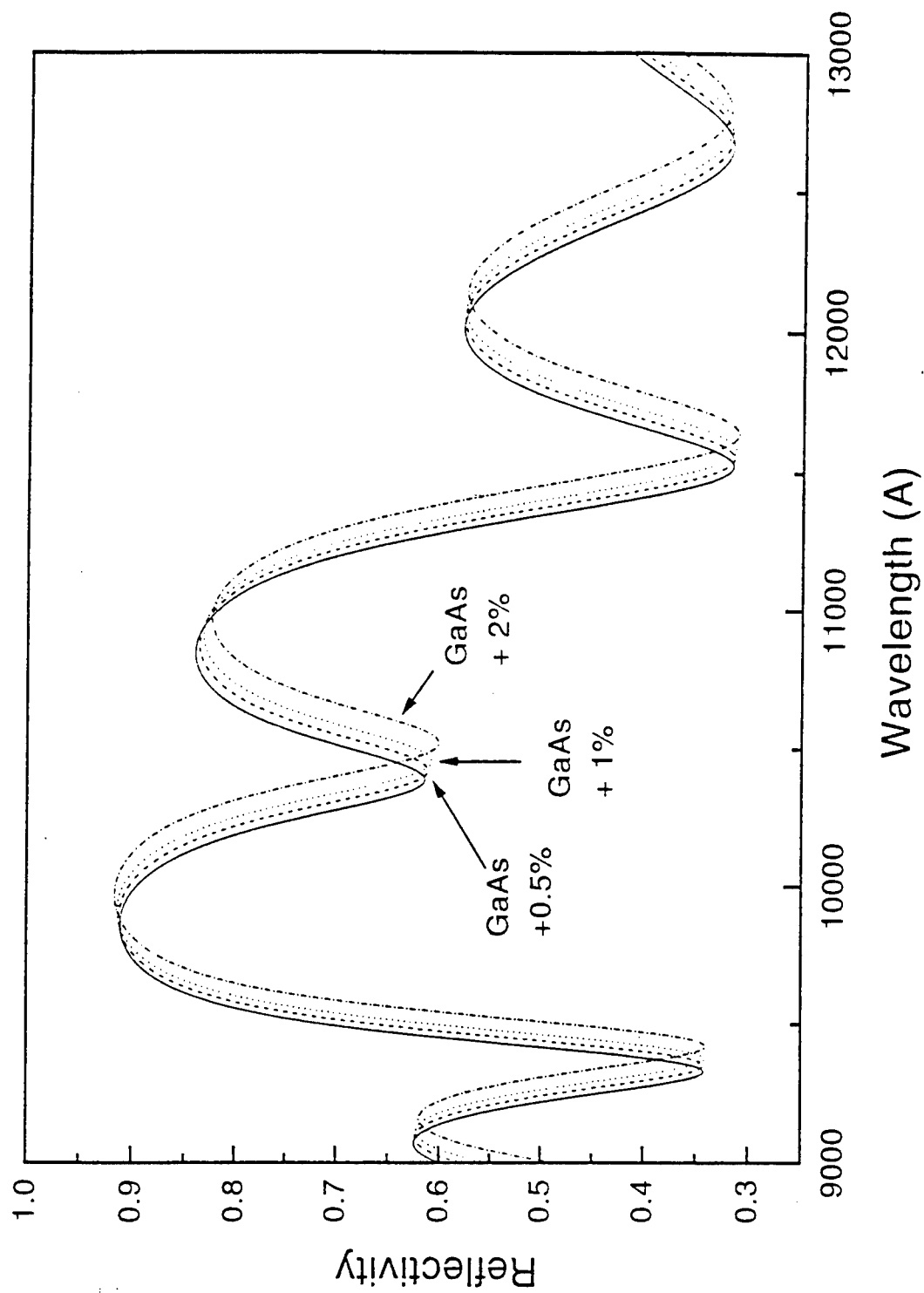


Fig.3. Simulated reflectivity spectra for the multilayer AlAs/GaAs stack with four different GaAs layer thickness. The dashed lines indicate the effect of increasing the GaAs thickness by +0.5, +1, and +2%, respectively. [1]

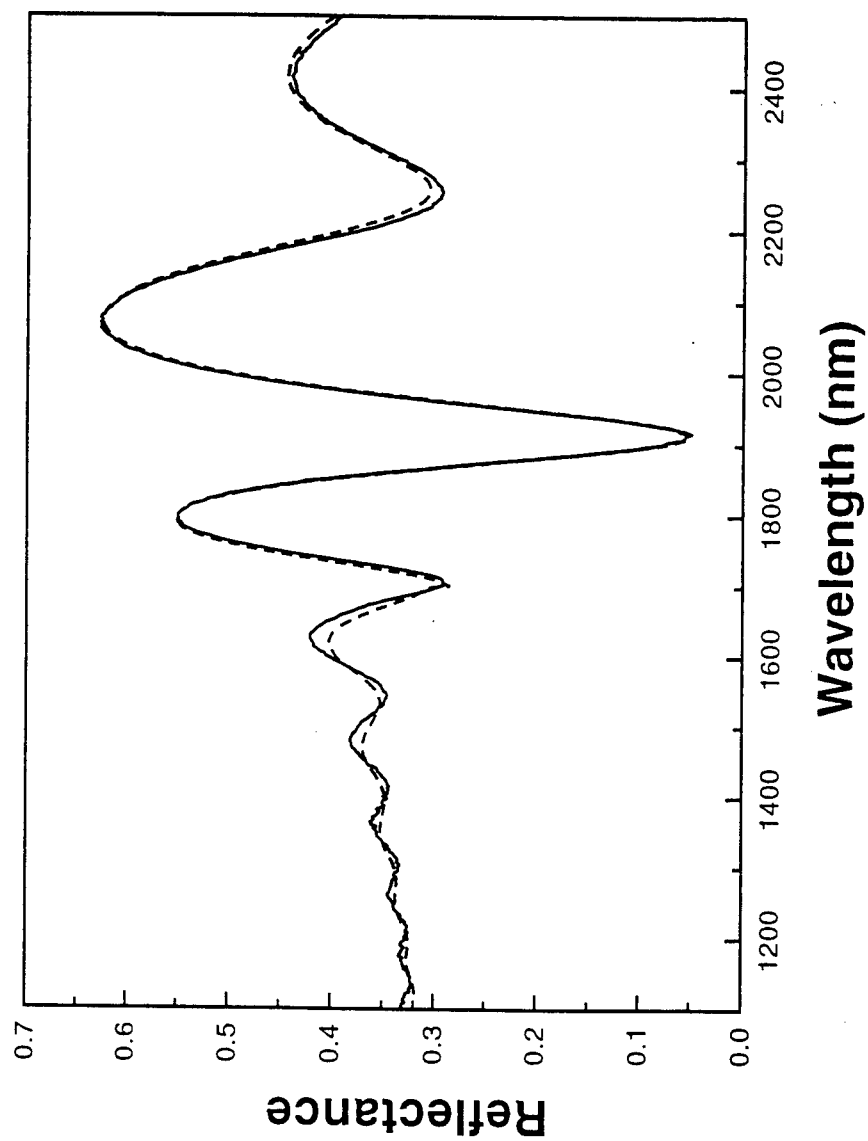


Fig.4. Experimental (solid) and theoretical (dash) reflectance spectra from a 7-pair $\text{In}_x\text{Ga}_{1-x}\text{As}/\text{InP}$ stack.

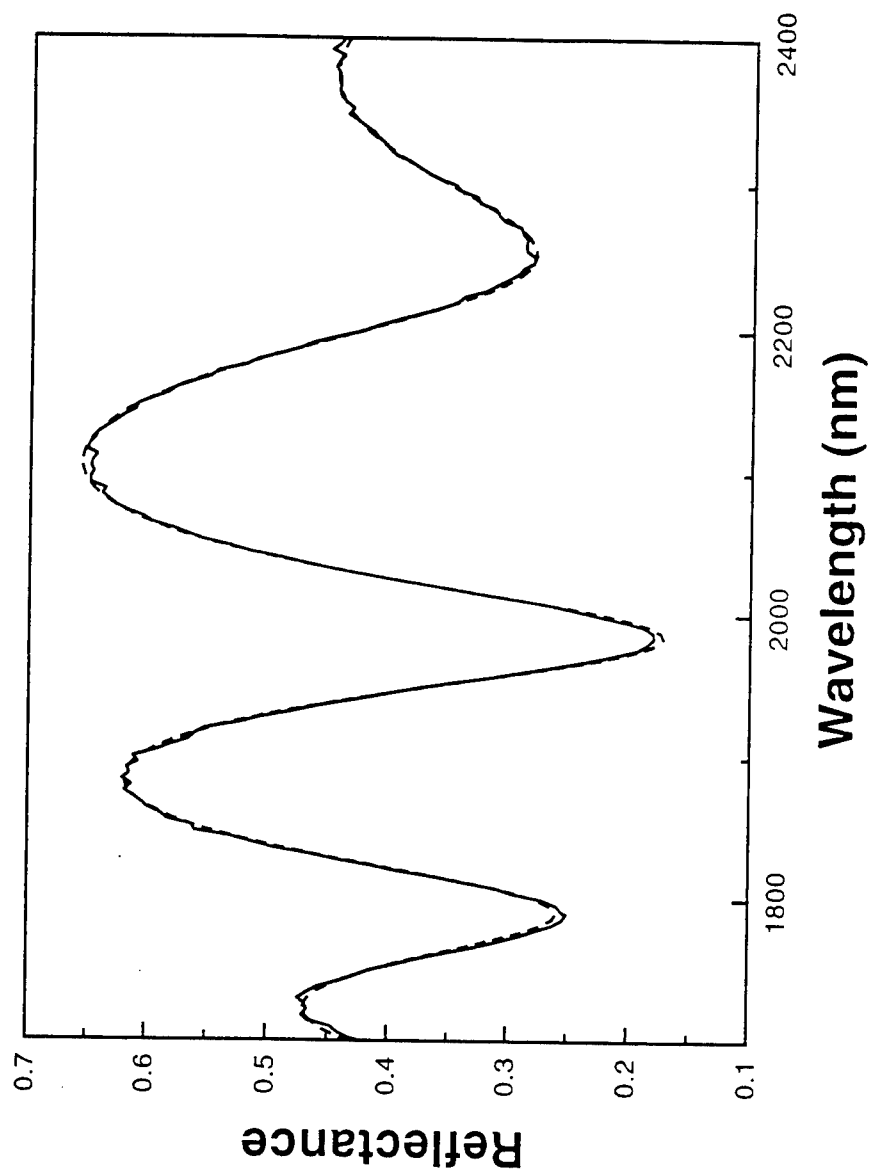


Fig.5. Experimental (solid) and theoretical (dash) reflectance spectra from a 9-pair $\text{In}_x\text{Ga}_{1-x}\text{As}/\text{InP}$ stack.

$\text{In}_x\text{Ga}_{1-x}\text{As}$ Composition Map

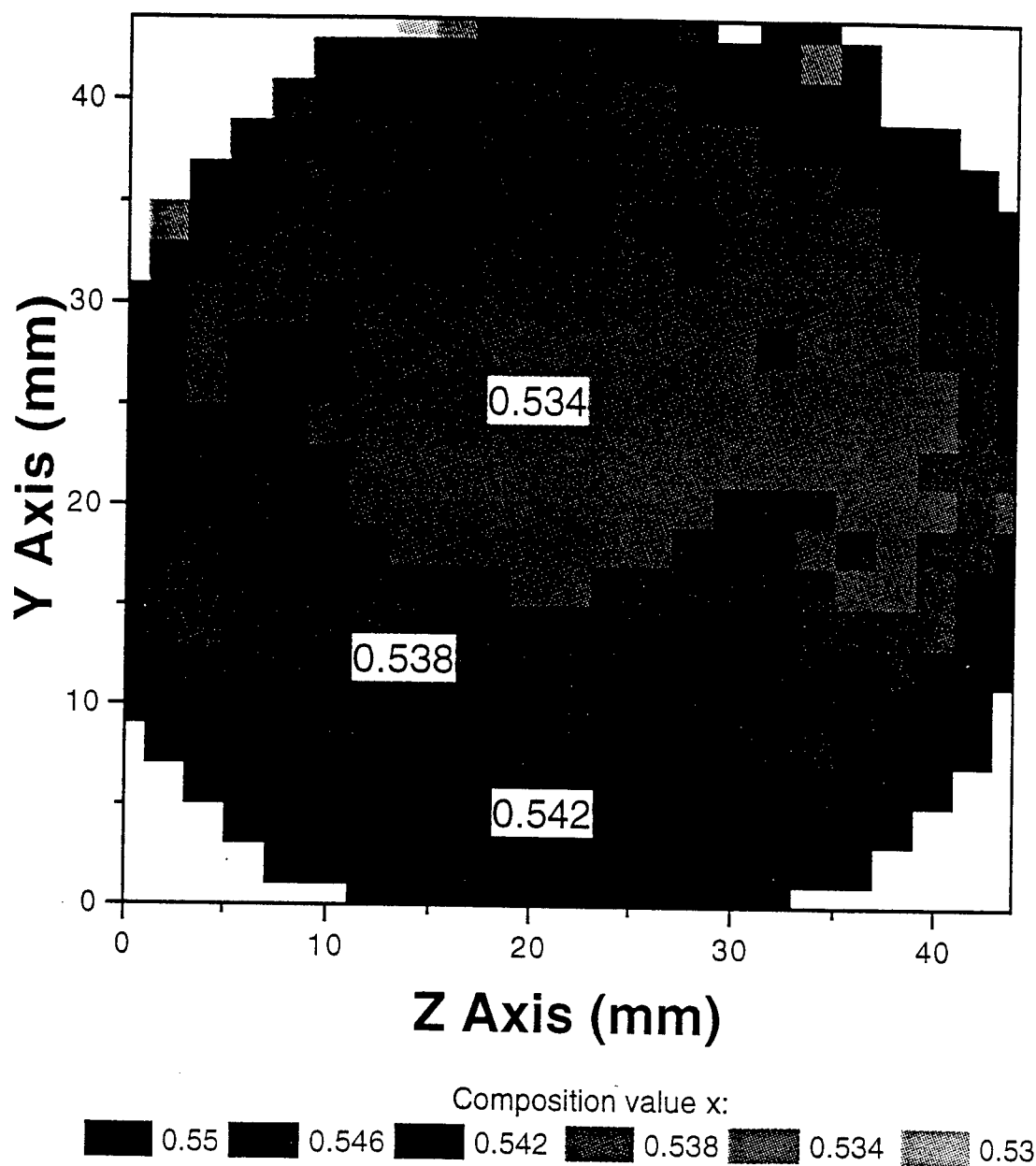


Fig.6. Composition (x) map for a 7-pair $\text{In}_x\text{Ga}_{1-x}\text{As}/\text{InP}$ wafer.

$\text{In}_x\text{Ga}_{1-x}\text{As}$ Layer Thickness Map

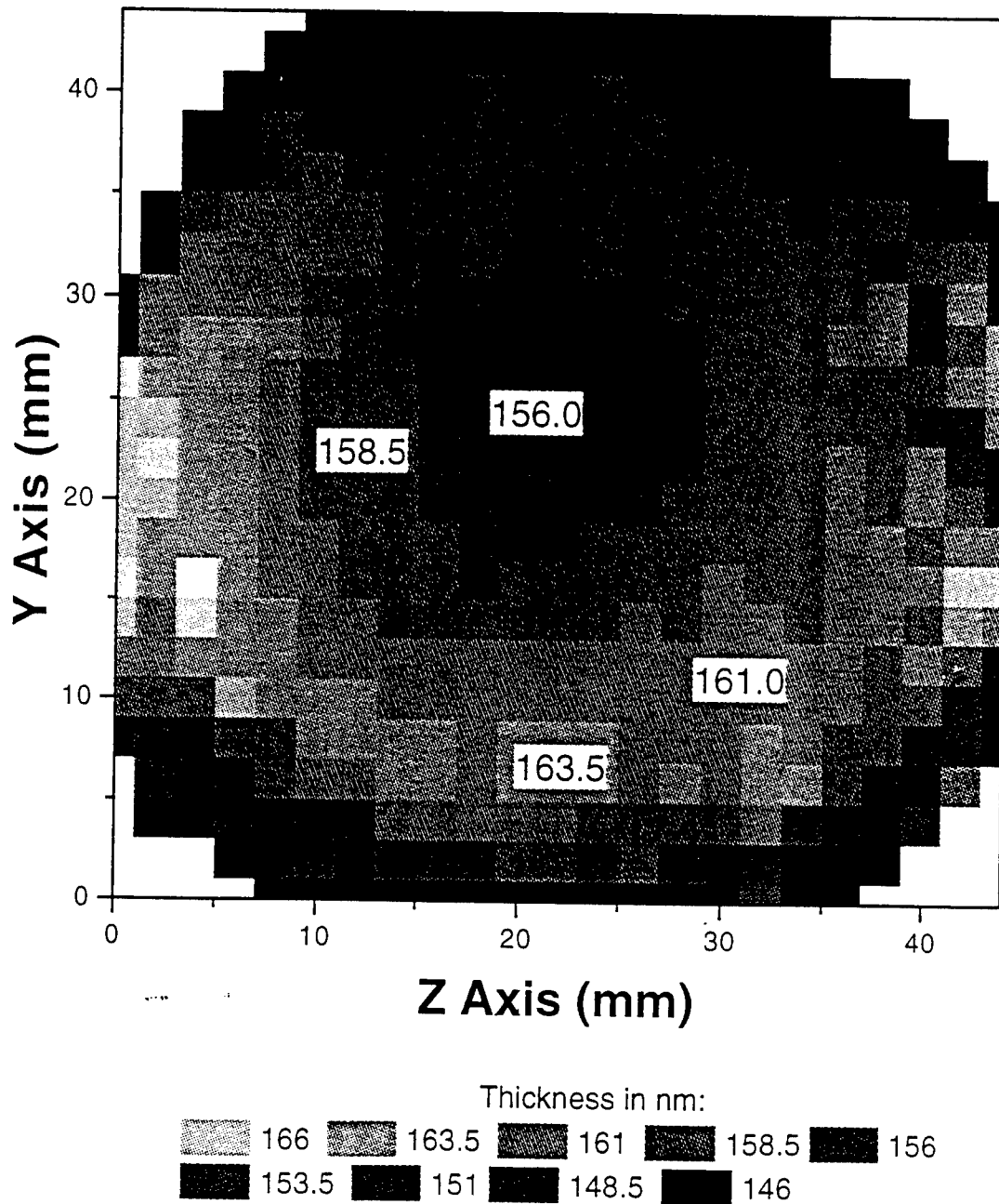


Fig.7. $\text{In}_x\text{Ga}_{1-x}\text{As}$ thickness map for a 7-pair $\text{In}_x\text{Ga}_{1-x}\text{As}/\text{InP}$ wafer.

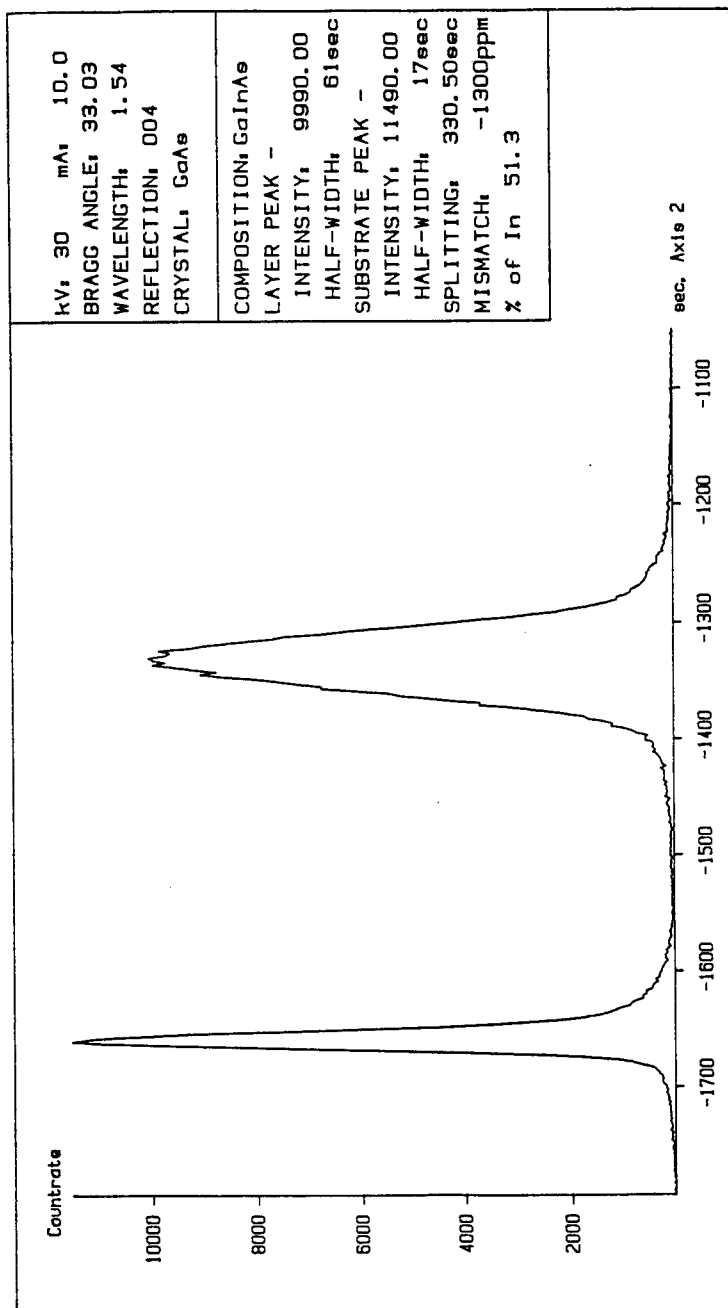


Fig.8. Double crystal X-ray diffraction pattern from a single-layer $\text{In}_x\text{Ga}_{1-x}\text{As}$ on InP substrate.

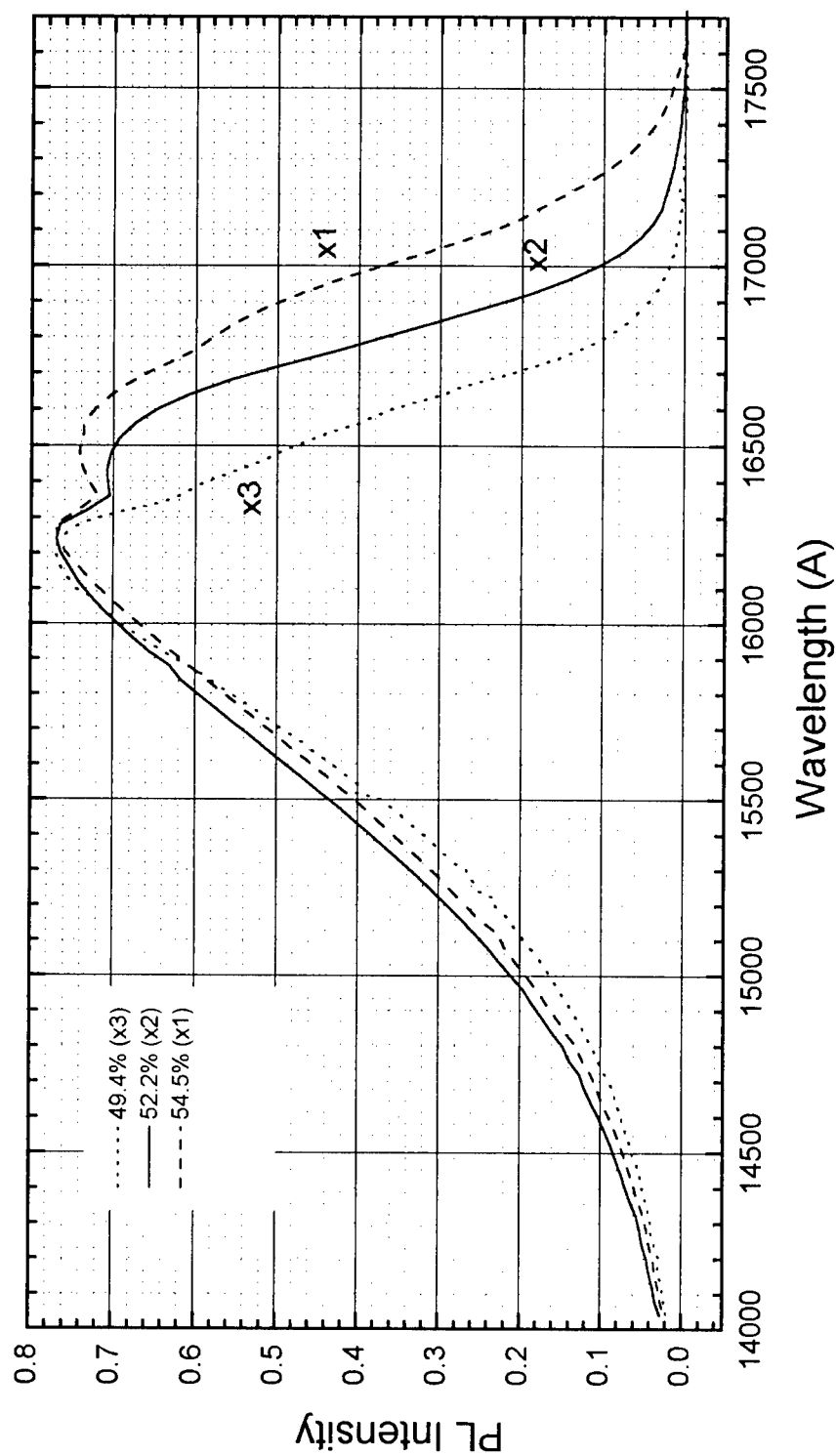


Fig.9. Photoluminescence spectra from 3 samples of $\text{InGa}_{1-x}\text{As}$ on InP substrate with different indium composition: $x1=54.5\%$, $x2=52.2\%$, and $x3=49.4\%$.

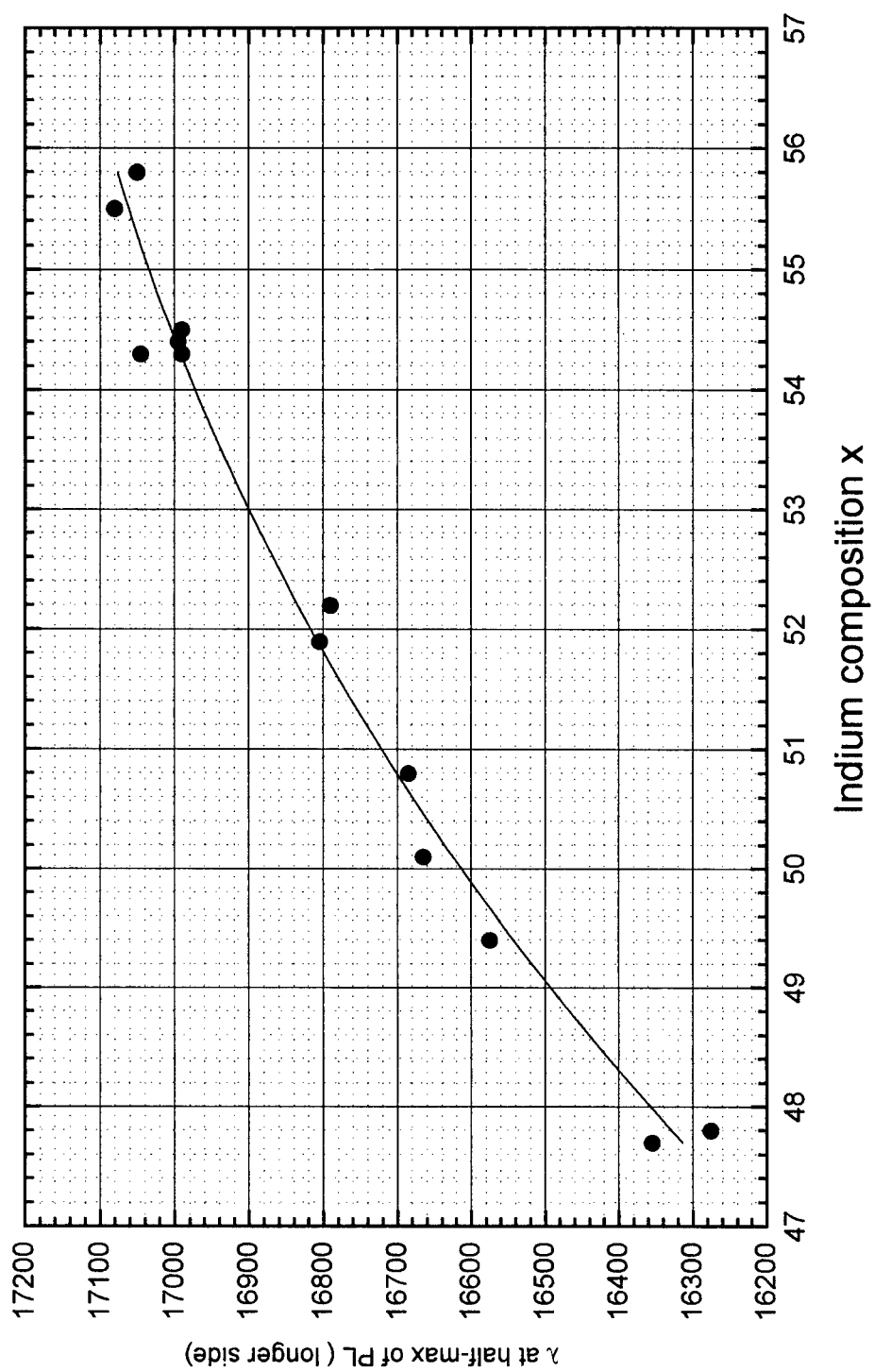


Fig.10. Wavelength position at half-max of PL on the longer side versus Indium composition for single-layer $\text{InGa}_{1-x}\text{As}$ on InP substrate.

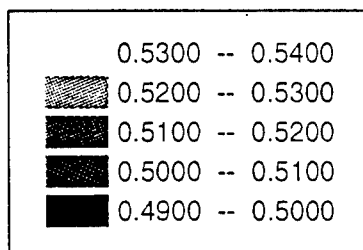
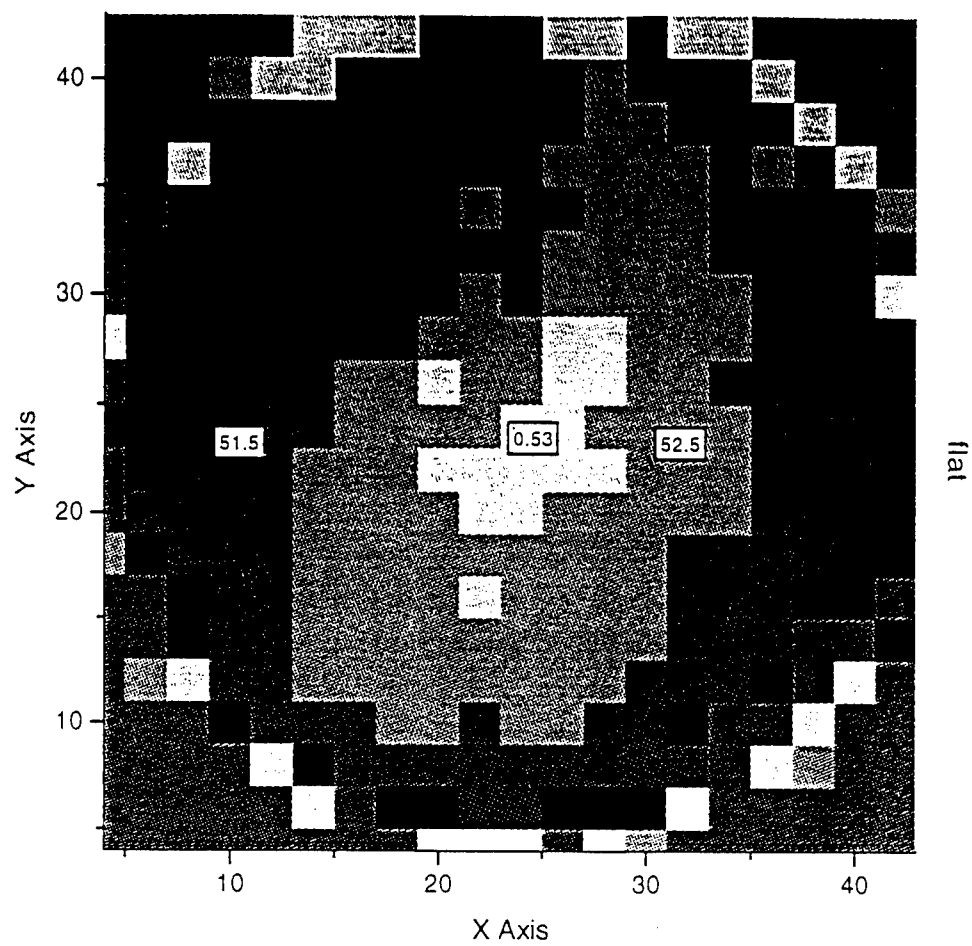


Fig.11. Composition (x) map using PL method for a 9-pair $\text{In}_x\text{Ga}_{1-x}\text{As}/\text{InP}$ stack.

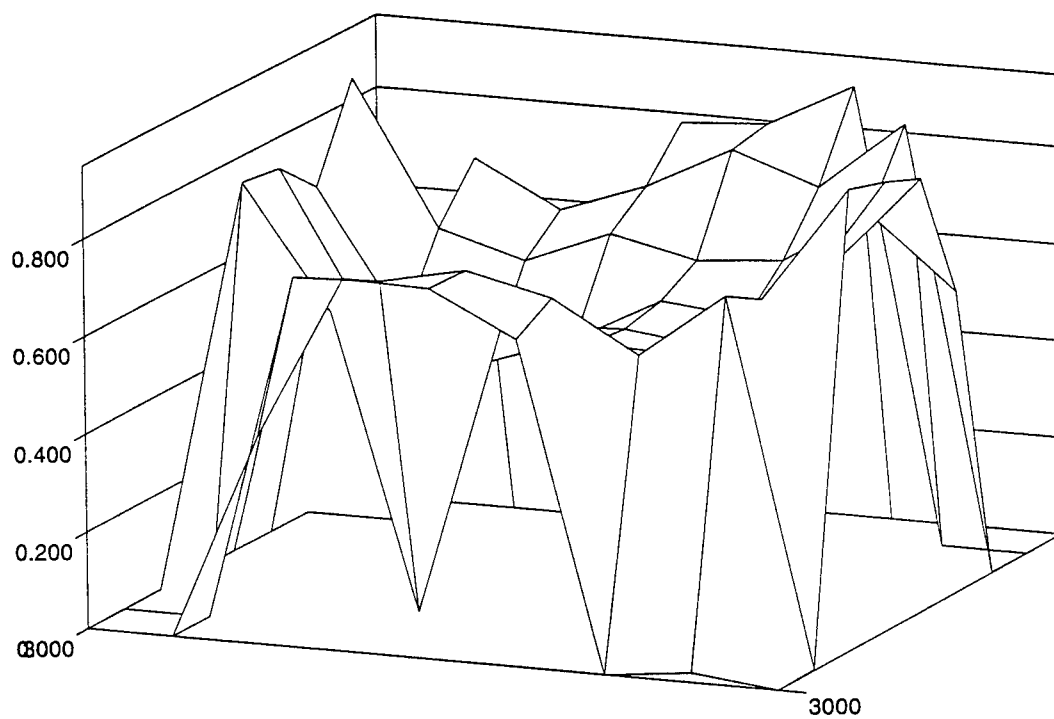


Fig.12. PL peak Intensity map for a 9-pair $\text{In}_x\text{Ga}_{1-x}\text{As}/\text{InP}$ stack.

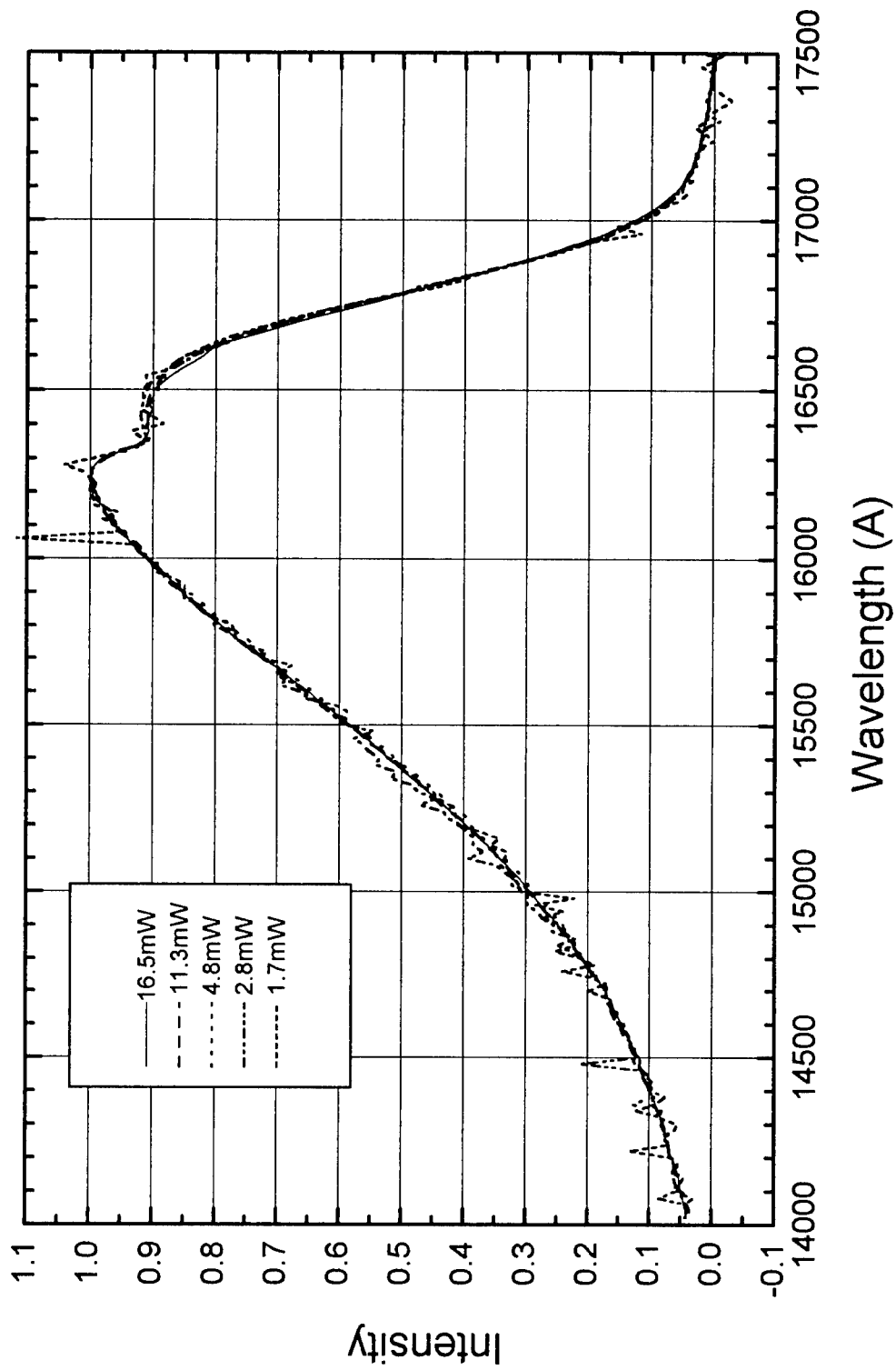


Fig.13. Photoluminescence spectra at different diode laser powers from a single-layer $\text{In}_x\text{Ga}_{1-x}\text{As}$ on InP substrate.

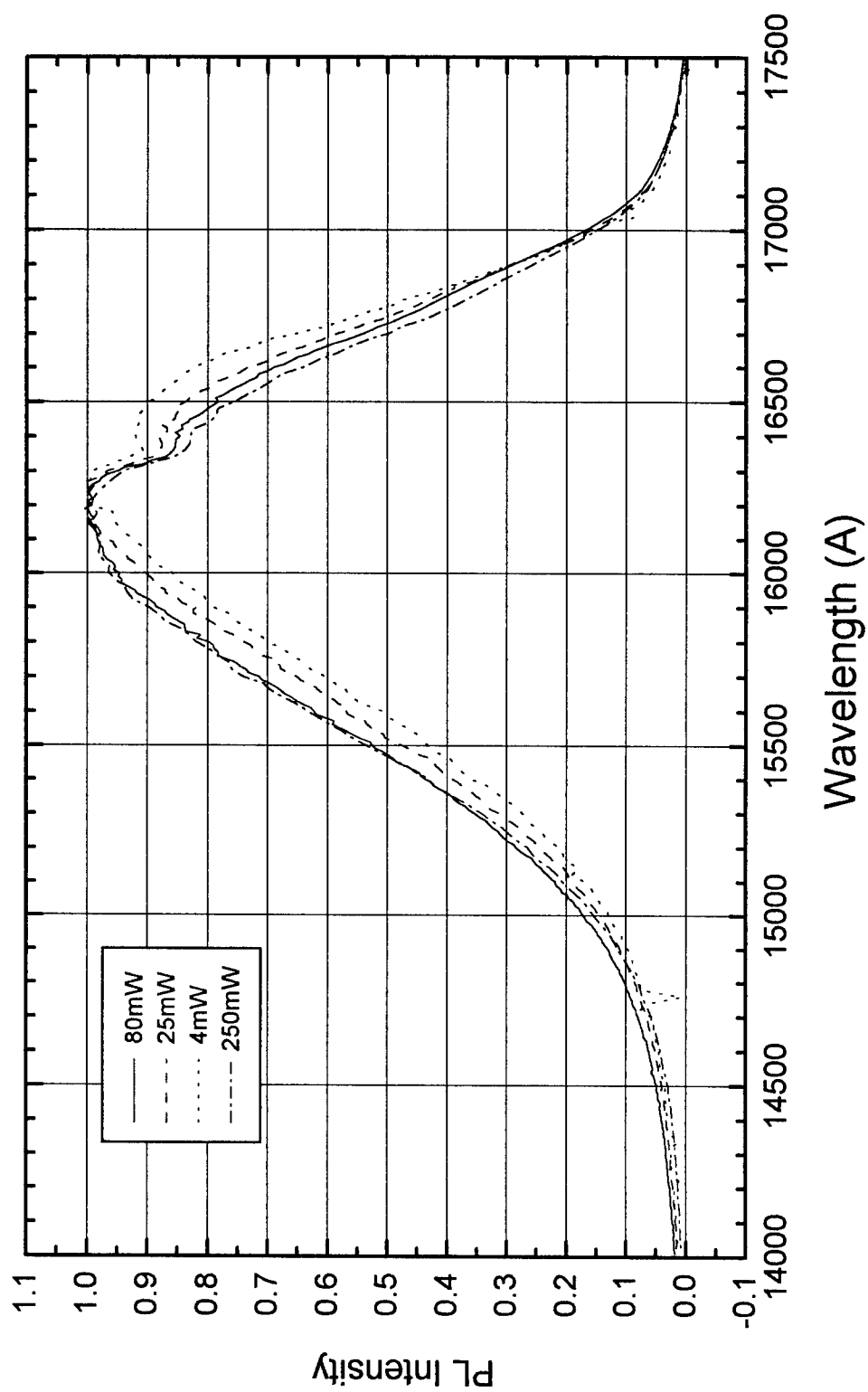


Fig.14. Photoluminescence spectra at different Argon laser powers from a single $\text{InGa}_{1-x}\text{As}$ on InP substrate.

Jun Chen
Report not available at time of publication.

**DEVELOPMENT OF ANTI-REFLECTION THIN FILMS FOR IMPROVED
COUPLING OF LASER ENERGY INTO LIGHT ACTIVATED ,
SEMICONDUCTOR RE-CONFIGURABLE , MICROWAVE
SOURCE/ANTENNA**

Prof. Everett E. Crisman
(Assoc. #96-0210)
Prof. Gang Xiao
Snorri Ingvarsson

Department of Physics
Brown University, Box 1843
Providence, Rhode Island 02912

Sponsored by:
Air Force Office of Scientific Research
Bolling AFB
Washington, DC
&
Brown University
Providence, Rhode Island

DECEMBER, 1997

ABSTRACT

We report the development and evaluation of a thin film anti-reflection coating for use on gallium arsenide that enhances the coupling of laser energy, at specific frequencies, into that material. The layers are of the single index, quarter wave type, matched to specific semiconductor specimens and to laser wavelengths used for optical excitation of various semiconductors. The purpose of this development is to increase the laser power coupling into the semiconductor specimens and thereby increase the radiated E-M field strengths of such elements when used for flyable, reconfigurable, μ -wave source/antenna arrays. This investigation compliments the ongoing project at AFOSR Sensors Technology Branch to develop semiconductor, E-M source/antenna elements for two and three dimensional radar arrays for airborne systems based on the concept.

INTRODUCTION

Optically excited semiconductor photo carriers, accelerated in a dc field, was suggested some years ago [1] as a potential source for wide band microwave pulses in the pico-second width range. Such sources would have a time domain width controlled (approximately) by the duration of the optical pulse and by the semiconductor photo carrier lifetime. Optical excitation permits multiple E-M source generation via splitting of a single laser beam. Phase and impedance match problems inherent in UHF microwave source - to - antenna coupling can be significantly reduced with this scheme. In addition, the E-M sources, as described, could act as their own radiative elements (antennae) thereby producing a compact array which can be steered electro-optically [2,3] rather than by complex mechanical and/or electronic delay lines. This would be a distinct advantage in high "G" environments. Cooling to at least LN₂ temperature is also feasible for airborne applications and has the potential for improving the field strength of such arrays by increasing the mobility and hence the final velocity attained by the optically induced, semiconductor carriers [6].

Research into various aspects of the laser induced, pulsed, picosecond, E-M sources (LIPPEs) has proceeded continuously, albeit at a low level, since 1994. The initial proof of concept was demonstrated by several organizations including Rome Laboratory, Hanscom MA [4]. The concept, simply stated, is that photo carriers, induced in a semiconductor by an optical laser pulse will accelerate in the presence of a dc. electric field established along the surface of the semiconductor (say between two surface metallic contacts on a thin polished wafer). Such accelerating carriers (generally electrons) will radiate electromagnetic fields in proportion to the applied dc. field strength up to some maximum velocity controlled by the intrinsic semiconductor parameters vis-a-vis the mobility. Experimental results of the past two years have generally confirmed this hypothesis. The general dc. field dependence has been examined and reported recently by the Liu, et al. at USAF Rome Lab. Hanscom, MA [5]. In that study, E-M radiation field strength GaAs and InP were examined as a function of applied dc. bias and it was demonstrated that, for both materials, a plateau in the radiated field was reached for the dc. field above some threshold – about 5.5 kV for GaAs and 12 kV for InP. Based those

studies and other information develop thus far, InGaAs should to be an even better source of E-M radiation producing stronger E-M fields at lower dc. voltages.

BACKGROUND

Recently, we completed a proof of principal program to evaluate the next level of complexity for the concept. As part of that evaluation [7], a single source was excited in two spatially separated regions by splitting the exciting laser beam. In an alternative configuration, two specimens, stacked in physically in series and biased separately, were excited simultaneously by the (split) laser beam. Near and far field measurements confirmed that a phase relationship could be superimposed on the detector by the opto-mechanical relationship between the source(s) and/or the exciting laser beam(s). That is, by varying the physical position on the surface and/or arrival time of the optical pulses, the forward direction of the maximum E-M position could be controlled.

The results of that study have suggested several courses that might be followed in order to increase the E-M signal strength. Eventually, the effort will devoted to optimizing the choice of semiconductor, the physical layout of the contacts, the profile for the dc. accelerating voltages and the spatial relationships between the semiconductor sources and the light excitation beam. One parameter that is of utmost importance to all the measurements, now and in the future, is the optical coupling between the light source and the semiconductor. This parameter presents challenges even now at this early stage of evaluation because of the nearly 35% losses at the air/semiconductor interface and the 50:50 power division between the two light beams from a single source. Therefore, for this study, we concentrated on the development of an anti-reflection coatings for one the two semiconductors that are presently the focus of the LIPPES studies.

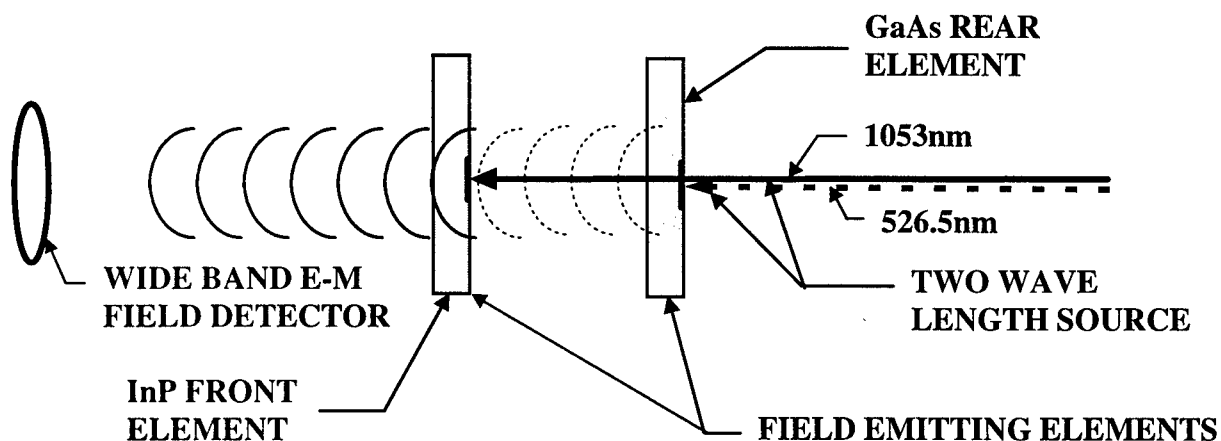


FIGURE 1: LIPPES series configuration for a two 'color' laser source exciting two E-M radiating element in series configuration.

For the experiments performed to date, the light source has been a Q switched, mode locked, frequency doubled, YLF laser with (ML/QS) power of 1.35 Watts and pulse energy of approximately 50μJ for 50ps duration pulse. The wavelengths available from the YLF laser are 1053nm and (frequency doubled) 526.5nm. Assuming laser light, at

normal incidence, impinging from an air medium onto a non-absorbing, partially reflective, polished surface then the reflectivity, for that single interface condition, is given by the Fresnel reflection equation:

$$R = \left[\frac{n_s - n_A}{n_s + n_A} \right]^2 \quad (1)$$

Where n_s is the index of refraction for the ambient medium (air, $n_A \approx 1$) and n_s is the index of the solid (semiconductor). Taking GaAs (the most used semiconductor of these studies thus far) as an example, then $n_s \approx 4.025$ and the reflection will be 36%! For InP, the other semiconductor use to date, the reflective losses are somewhat less at 31%. Finally, for InSb or InAs, which appear to be an even better E-M sources based on their mobility, reflection loss could be as high as 41%. It is obvious even from these simple estimations that dividing the laser source between many targets will significantly affect the semiconductor excitation level and hence the resulting E-M field strength which is the fundamental gauge for most of the evaluations thus far.¹ Therefore, even the basic development efforts underway now would profit by increasing the light coupled into the semiconductors and improving the precision of the research.

DISCUSSION OF RESEARCH PROBLEM

Anti-reflection coatings can range from a simple, single layer having virtually zero reflection at a particular wavelength to a variable index or multi layer system having minimal reflectance over a range of several octaves. The particular choice depends on a number of intrinsic parameters of the physical system, in particular, as already mentioned, the indices of refraction for both the coating and the substrate (assuming an air or vacuum ambient). These parameters are not as simple as often assumes and are, in fact, complex numbers the real and imaginary components of which are *wavelength dependent*. Thus the true index would be given by:

$$\eta = n - ik \quad 2)$$

where n is the (traditional) real index of refraction and k is the extinction coefficient. While k can usually be made quite small for most dielectric, it is in general finite and of the same order as n in semiconductors. It can, therefore, not be ignored. n and k are related to the particular solid through the high frequency dielectric constant which itself is a function of the light wavelength.

$$n^2 - k^2 = \epsilon(\lambda) \quad (3)$$

The optical constants, n and k , are both strongly varying functions of the light frequency, $\nu = 1/\lambda$, particularly at wavelengths near the absorption edge of the

¹ The numbers calculated are based on published information for the indicated semiconductors [7,8]. and assume a visible green wave length of 526.5nm for the YLF laser source after frequency doubling. Some of the variations suggested for improving the E-M field strength will require AR coatings optimized for $\lambda \approx 1053\text{nm}$ and other wavelengths near 1000nm.

semiconductor. Other effects such as impurity levels crystallographic imperfections and surface preparation (damage) play significant roles in the effective optical constants of a particular specimen. Finally, the polarization of the light itself is an important factor in determining R. Generally the LIPPES experiments maintain 'p' polarization to minimize the reflection at the beam splitters, dielectric beam steering mirrors and semiconductor surface. For reasons related to the E-M radiation pattern in the LIPPES experiments, normal incidence is preferred. In that case the actual reflectivity equation becomes:

$$R = \frac{(n_s - n_A)^2 + k_s^2}{(n_s + n_A)^2 + k_s^2} \quad (4)$$

Since n_s and k_s can be analytically derived from the parameters measured by research ellipsometers, those values will be the starting point for the design of the AR coatings.

Once the optical constants of the particular semiconductor wafer surface are known the task of determining the AR coating depends primarily on the wavelength range for which reflection must be a minimized. In the case of the very narrow bands represented by laser sources, single AR layers of *single index* can usually be defined which will reduce total reflection to a few percent. As this will be a significant improvement for the LIPPES measurements, was the direction chosen for this program. Assuming then a single layer, it is readily shown that the condition on the index of the coating, n_c , for the minimum reflection is:

$$n_c = (n_A n_s)^{1/2} \quad (5)$$

provided that the coating has an optical thickness one quarter of the light source wave length, i.e.:

$$n_c d = \lambda/4 \quad (6)$$

where d is the actual thickness.

(At normal incidence the correction for absorption coefficient can, and will, be neglected. The challenge then, for GaAs, (with $n_s \approx 4$ at 1056nm) was to specify and deposit a dielectric would have an index in the variable about 2.00 ± 0.10 , that will withstand the humidity and common solvent fumes of the laboratory and which can be formed into sputtering targets from available high purity chemicals.² For example, using GaAs with $n_s \approx 4$ at 527nm, the simple calculation of From various published sources some AR candidates can then be defined. For example, Si_3N_4 has $n = 1.97$ to 2.02 when deposited by sputtering and readily available in very high purity. There is now tabulated

²It should be noted that the added flexibility provided by double layer AR coating can almost certainly provide $R \approx 0$ for single wavelength sources at normal incidence [8]. Since double layering allows both thickness and index of the AR films to be varied, it is possible to choose for the interface AR layer a surface passivating compound. Surface passivation will allow the dc. bias field to be increased as necessary to insure velocity of saturation of the accelerated charges. So, this research also provides an excellent foundation for future efforts to optimize the LIPPES technique for airborne and other applications.

information on a variety of dielectric compounds for which the refractive index has been measured at least over some portion of the visible and near infrared spectrum [9]. It is also possible to 'tailor' a coating for a specific index values. Ternary mixtures such as silicon oxynitride, SiO_xN_y , with $1.46 (\text{SiO}_2) \leq n_c \leq 2.10 (\text{Si}_3\text{N}_4)$ and aluminum oxynitride, AlO_xN_y with $1.76 (\text{Al}_2\text{O}_3) \leq n_c \leq 2.2 (\text{AlN})$ would provide a composition tunable index of high durability (hardness and moisture resistance).

EXPERIMENTAL APPROACH

AR Coating Preparation:

All films were prepared by reactive RF magnetron sputtering in a SS chamber pre cryo-pumped (15°K , closed cycle He) to $P < 1 \times 10^{-7}$ Torr. After the specimens were introduced into the vacuum chamber and prior to introduction of the sputtering gases, the system was baked to $T > 100^\circ\text{C}$ to reduce the residual oxygen partial pressure. Approximately 12hrs was required for the bake out cycle. Zero grade argon and nitrogen were introduced through PID gas flow controller and associated valves. Total pressure for the film production was 5mT for the combined Ar+N₂ gases (typically 4mT Ar and 1mT N₂ during the sputtering cycle). The target for all the runs of this study were polished, N-type semiconductor grade, silicon wafers doped to 1×10^{16} with phosphorous. The specimens were nominally cleaned by rinsing in acetone and/or methanol and blown dry with N₂ before loading into the UHV chamber. The target to specimen distance was 18cm. All runs were done with total RF power at 200 Watts.

In order to develop a baseline for sputtered Si_3N_4 thickness, calibration runs were done with silicon wafers (rather than GaAs) as the substrate specimens. Undoped silicon wafers were also used as the sputtering target and there was some danger in sputtering through those damaging the sputtering gun. Therefore, the depth of the sputtering crater was closely monitored during the calibration runs to track target oblation. Typically three runs were done before the silicon *target* was changed.

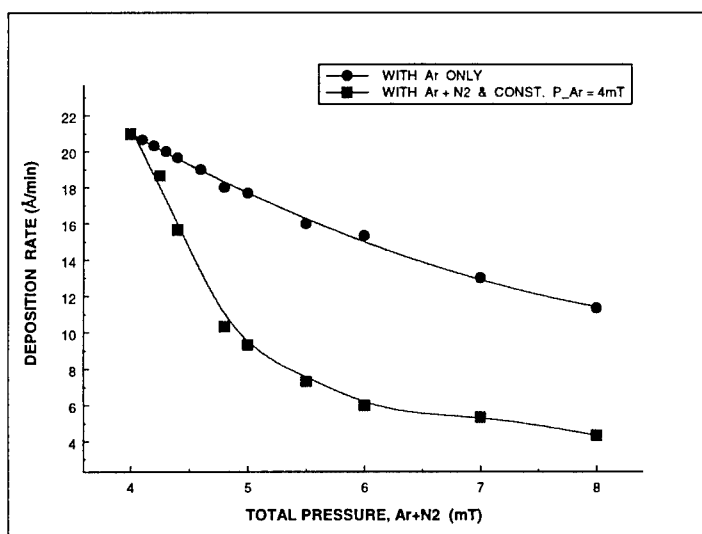


FIGURE 2:
Comparison of reflectivity of GaAs measured for these studies with reported data from the literature⁸.

The data in figure 2 shows the effect of the N₂ partial pressure on the Si_xN_y deposition rate. As predicted there is an inverse relationship of deposited rate with increasing partial pressure of N₂ (in Ar).

Based on the information developed above the following films were prepared to specifications provided to us by the Air Force scientists at Hanscom AFB:

| Specimen number | Design thickness (Å) | Sputtering time (min) | Measured thickness (Å) | Index of refraction, n _A |
|-----------------|----------------------|-----------------------|------------------------|-------------------------------------|
| 5239 | not specified | 40.0 | 1670 * | * |
| 5240 | 938 | 20.0 | 945 | 1.973 |
| 5241a | 656 | 15.5 | 693.3 ± 2.6 | 1.947 ± .001 |
| 5241b | 649 | 16.0 | 627.8 ± 3.4 | 1.966 ± .001 |
| 5242 | 976 | 22.5 | 969.4 ± 3.3 | 1.956 ± .002 |

*NOTE: Ellipsometry on specimen 5239 gave undetermined results thus there is no measurement for n. The thickness value is from Dektak measurements. From this thickness an approximate sputtering time of 20 min. was calculated for the next run.

Reflectivity Measurements:

Evaluation of the Si₃N₄ films as AR coatings was done performed by measuring the reflectivity directly as a function of wavelength using a LEXEL, model 480 Avante, Ti- Sapphire laser excited by a 5 Watt, CW SPECTRA PHYSICS, Millennia, diode pumped, YAG laser (frequency doubled to 532nm). The LEXEL was fitted with dielectric cavity mirrors optimized for a wavelength range of 700 <λ< 850 nm and was wavelength tunable via a rotatable bi-refracting crystal. A dielectric mirror reflected 5% of the extant light into a NEW FOCUS, Inc. Model 7711 Fuzeau wavelength meter to monitor the λ parameter. Approximately 8% of the remaining power was reflected, with a microscope slide, to the detector element of a OPHIR Laser Power Meter used to monitor the incident power.³ The remaining light passed through the slide and, after reflecting from the specimen under test, reflected to a MOLECTRON EPM1000 energy/power meter fitted with a J3 detector. A program was developed using HP-VEE to acquire data from the individual measuring devices.

The procedure for obtaining the data was to adjust the bi-refracting crystal for a specific wavelength and then run the program which took ten readings at each λ setting. The data was stored in tabular form and later averaged to provide incident power, reflected power and reflectivity ($P_{\text{reflected}}/P_{\text{incident}}$) as a function of λ for the plots below.

Before each specimen was measured, a first surface mirror was placed at the specimen position, the two power meters were 'zeroed' with the laser blocked and the room lights off, and then the laser was turned on and the ratio of incident to reflected power was measured. This number was then use as the '100% reflectivity' value for adjusting the subsequent measurements with the specimen in place.

³ The microscope slide had essentially 'flat' transmission over the λ range used here and so no correction was made for its presence.

RESULTS

Before the specimens produced in the facilities at Brown University were evaluated the reflectivity setup, described above, was used to measure a specimen of un-coated GaAs of the type that was intended for AR coating. This was done after successively cleaning in e-grade acetone, e-grade methanol and etching in HF:NH₃OH, rinsing in DI-water and finally blow drying in dried zero grade nitrogen. The specimen was measured as described above and then the plot of reflectivity versus wavelength was generated as shown in figure 3.

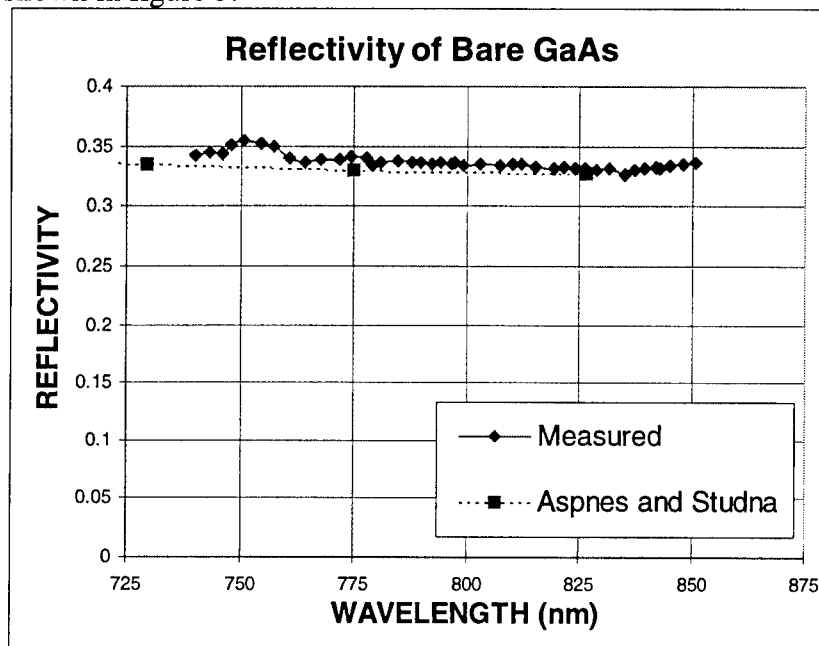


FIGURE 3:
Comparison of
reflectivity of
GaAs measured
for these studies
with reported
data from the
literature⁸.

Added to that plot is the data of Aspnes and Studa⁸ which was made on vacuum cleaved GaAs at 1×10^{-9} T and is the most accurate available for GaAs. As can be seen in the figure, the data for this study tracks well the published data except for small variations in the region around 750nm. Based on the good agreement here the system was applied to the AR coated specimen in the same fashion.

The plot of figure 4 shows the reflectivity of the 167nm coated specimen(#5239). While the 167nm AR coating was prepared primarily as thick layer on the first GaAs specimen for reference purposes. When it was measured for reflectivity, it was found to approach a minimum at in the vicinity of 760nm. This is in agreement with the predicted minimum λ for a film of 167nm thickness which should occur at an odd multiple of $1/4\lambda_c$, i.e. the wavelength corrected for the index of the AR coating. (In this case $N=17$ would imply a minimum for 161nm thickness). As seen in the figure the reflectivity not only is approaching a minimum but also has an extremely low absolute reflectivity on the order of 0.5%! This means that thicker layer are acceptable and might even have some advantages over the first quarter λ thickness due to better adhesion. There are two separate series of data plotted in figure 4.

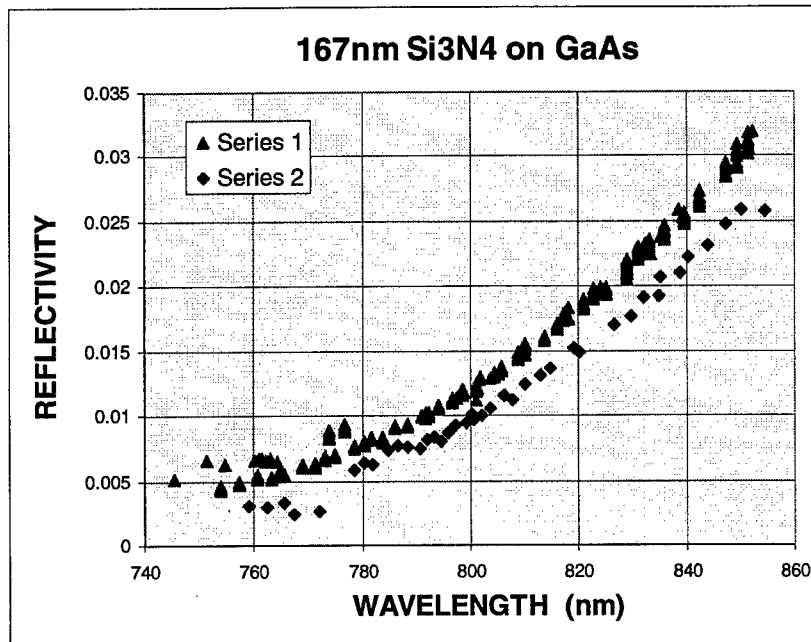


FIGURE 4:
Wavelength
dependence of
Si₃N₄ films
reactively
sputtered on to
GaAs specimen
#5239.

Series 1 was done with the small amount of 532nm light that bleeds through the optical path and originates from the diode/YAG laser used to excite the LEXEL Ti-Sapphire system. The series 2 data was a re-run of the same #5239 specimen done after filters had been added to remove that 532nm component. As can be seen, the reflectivity is further reduced when the stray background component is removed.

Of the three remaining specimens listed in the table above, #5241a was not measured because the film had clearly delaminated for the GaAs substrate leaving a mottled surface appearance.

AR coating 5421b had a design thickness of 64.9 nm chosen to minimize reflection at 526nm which is $\lambda/2$ of one of the laser lines intended for use as in the LIPPES system.

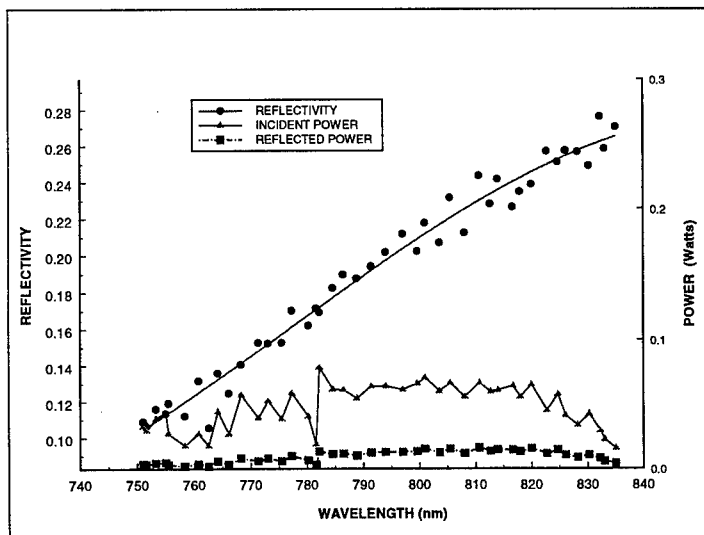


Figure 5:
Reflectivity versus
wavelength for
specimen #2541b
with 64nm AR
coating. Specimen
optimized for
minimum λ
reflectivity at 526 nm.

The incident and reflected power data and the resultant reflectivity are plotted for that specimen in figure 5. A polynomial curve fit has been added for the reflectivity

data. Extrapolation of that equation to shorter wavelengths give a minimum reflectivity at 670nm which is about a 21% error but with an absolute reflectivity on the order of 5%.

A similar plot is shown in figure 6 for AR coating #5242 with a 97.6 nm design thickness. Here again the original incident and reflected powers are plotted along with the resulting ratio (P_i/P_r) to give the reflectivity. The polynomial curve fit for that curve has

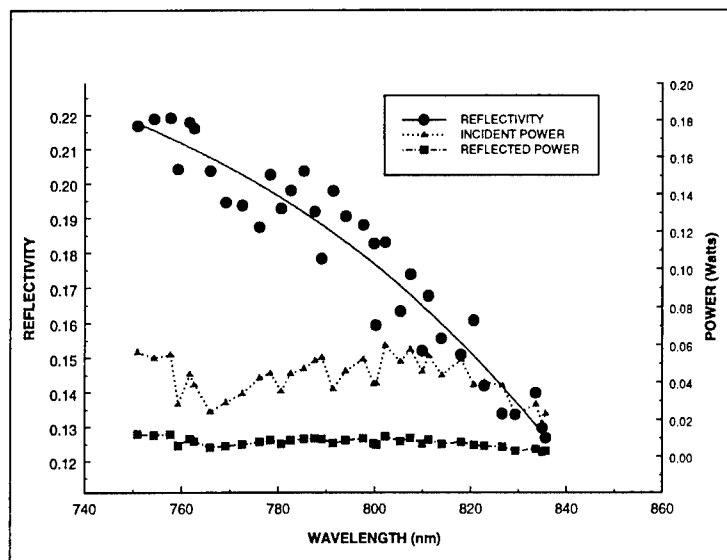


Figure 6:
Reflectivity versus
wavelength for
specimen #5242
with 97.6nm AR
coating. Specimen
optimized for
minimum λ
reflectivity at 800
nm.

appears to be approaching a minimum in the vicinity of 850 nm which implies an error on the order of 6%. Since leveling off of the curve is not observed here the error is likely greater but the reflectivity at minimum will clearly be less than 10% at the minimum. The main source of disagreement between the predicted minima and that indicated by the data is probably the relatively large area covered by the reflectivity probe beam. This allows the sampling of areas where there are gradients in the thickness and hence strong variations in the reflectivity values. Time constraints did not permit the investigation of that explanation, but experiments are continuing at the AF laboratory to evaluate that effect. In general we have shown that Si_3N_4 reactively sputtered films can be varied in thickness to provide reasonable matches to specific wavelengths. Refinements of this process should provide repeatability to within a few percent in λ with minima in reflectivity losses less than 5%. Broader band light sources will allow a more accurate determination of the values at the reflectivity minima.

CONCLUSION

We have developed a procedure for reactively sputtering continuous Si_3N_4 thin film anti reflection coating onto GaAs for the purpose of enhancing the transmission of certain discrete laser wavelengths. Independent measurements of absolute reflectivity show that those AR coatings reduce the reflective losses to $< 0.05\%$ and that the process control is sufficiently precise to allow the film thickness to be controlled to within $\pm 10\text{nm}$. This technology was transferred to the scientists at the AFOSR Sensors Technology Division who were able to reproduce the film characteristics and are adapting the process for their particular specimens, substrates and apparatus.

REFERENCES

1. Ch. Fattering and D. Grischkowsky, **Tetrahertz Beams**, *Appl. Phys. Lett.*, Vol. 54, pp 490 - 492, (1989).
2. B .B. Hu a. t. Darrow, X. -C. Zhang and D. H. Auston, **Optically Steerable Photoconducting Antennas**, *Appl. Phys. Lett.*, Vol 56, No. 10, pp. 886 - 888, (1991).
3. X. -C. Zhang and D. H. Auston, **Generation of Steerable Submillimeter Waves from Semiconductor Surfaces by Spatial Light Modulators**, *Appl. Phys. Lett.*, Vol. 59, pp. 768- -770 (1990)
4. D. W. Liu, J. B. Thaxter and D. F. Bliss, **Gigahertz Planar Photoconducting Antenna Activated by Picosecond Optical Pulses**, *Optics Letters*, Vol. 15, No. 14, pp. 1544 - 1546 (July, 1995).
5. W. Liu, P. H. Carr and J. B. Thaxter, **Nonlinear Photoconductivity Characteristics of Antenna Activated by 80 - Picosecond Optical Pulses**, *IEEE Photonics Technology Letters*, Vol. 8, No. 6 (June, 1996).
6. E. E. Crisman, **Evaluation Of Semiconductor Configurations As Sources For Optically Induced Microwave Pulses**, Final Report to AFOSR 1996 Summer Faculty Research Program, to be published (Dec., 1996)
7. D. W. Liu, E. E. Crisman, J. S. Derov, P. H. Carr and S. D. Mittleman, **Two and Three Dimensional Reconfigurable Arrays Using Optical Generation in Semiconductors as the Source -Antenna Elements**, presented at *The 7th Annual DARPA Symposium on Photonic Systems for Antenna Applications (PSAA-7)*, Monterey, CA, (January, 1997)
8. D. E. Aspnes and A. A. Studna, **Dielectric Functions and Optical Parameters of Si, Ge, GaP, GaAs, GaSb, InP, InAs, and InSb from 1.5 to 6.0 eV**, *Phys. Rev. B*, Vol 27, No. 2, pp. 985-1007, (1983).
9. **THIN FILM OPTICAL FILTERS**, by H. A. Macleod, Macmillian, New York NY, 2nd ed., pp 78 - 86 (1986).
10. See for example, **CRC Handbook of Laser Science and Technology, Supplement 2: Optical Materials**, M. J. Weber editor, CRC Press, pp. 30 - 64. (1995).

**DEVELOPMENT OF A SIMULATION MODEL
FOR DETERMINING THE PRECISION OF RELIABILITY
PREDICTIONS AND ASSESSMENTS**

**Digen K. Das
Associate Professor
Department of Mechanical Engineering Technology**

**SUNY Institute of Technology
P.O. Box 3050
Utica, NY 13504-3050**

**Final Report for:
Summer Research Extension Program**

**Sponsored by:
Air Force Office of Scientific Research
Bolling Air Force Base, Washington DC**

and

SUNY Institute of Technology at Utica/Rome, Utica, NY

June 1998

DEVELOPMENT OF A SIMULATION MODEL
FOR DETERMINING THE PRECISION OF RELIABILITY
PREDICTIONS AND ASSESSMENTS

Digen K. Das
Associate Professor
Department of Mechanical Engineering Technology
SUNY Institute of Technology

Abstract

Simulation models for determining the precision of reliability predictions for series, parallel and combination systems were studied by applying both the probability and possibility theories. The method of bounds and fuzzy number technique were used to determine the precision of the reliability predications. Available field data for capacitors (commercial and military) were used for all computational results.

DEVELOPMENT OF A SIMULATION MODEL FOR DETERMINING THE PRECISION OF RELIABILITY PREDICTIONS AND ASSESSMENTS

Digen K. Das

Introduction

It is well known that there exists a significant difference between predicted reliability and actual field reliability of systems. This is often termed as "Reliability Delta" and has been a long standing problem in reliability engineering. Usually system reliability is derived from the knowledge of component reliability and the concepts of statistical life-time distributions. However, in practice this knowledge does not necessarily provide insight into how a single component will behave in a system. As a consequence, systems all too often achieve reliabilities markedly different (usually lower) than those predicted.

In response to this long standing problem, the Air Force Research Lab initiated the "New System Reliability Assessment Methods" program (Ref. 1, 2). The objective of this effort was to develop a system reliability methodology that accounts for all predominant factors that affect field reliability of systems. The performing organizations in this program were IIT Research Institute/Reliability Analysis Center (IITRI/RAC) and Performance Technology.

The Summer Faculty Research Program (Ref. 3) was designed to be a complementary effort to the on-going "New System Reliability Assessment Methods" program. The objective of the Summer Faculty Research Program was to initiate the development of practical methods for determining the accuracy of reliability predictions and assessments. It was established that the models for determining the precision of reliability should be based on both the probability and

possibility theories. This Summer Research Extension Program addresses this research topic.

The organizational structure of most systems can be described as a series system, a parallel system or a combination thereof. The mathematical models for determining the reliability of these systems are described in the following sections.

Mathematical Models [Tasks 1 and 2]

Series System

The mathematical reliability model for a simple series system is the product of the individual reliabilities (Ref. 4). Figure 1 depicts such a series system with n elements. It is assumed that each element (capacitor) is independent—that the success or failure of one element does not affect the success or failure of any other element.



Fig. 1 Logic diagram of a simple series system of n elements.

The model is given by the following equation.

$$R_T = R_1 \times R_2 \times R_3 \times \dots \times R_i \times \dots \times R_n \quad (1)$$

where R_T = the total reliability
 R_1 = the reliability of the first unit
 R_i = the reliability of the i th unit
 R_n = the reliability of the n th (last) unit

This can be written in a more condensed form

$$R_T = \prod_{i=1}^n R_i \quad (1a)$$

where $\prod_{i=1}^n R_i$ means the product of all R_i from 1 through n .

Here R_i can be calculated from the following relation:

$$R_i = e^{-\lambda_i t_i} = e^{-F_i}$$

where λ_i = constant failure rate
 t_i = mission operating time
 $F_i = \lambda_i t_i$

Therefore,

$$\begin{aligned} R_1 \times R_2 \times R_3 \times \dots \times R_i \times \dots \times R_n \\ = e^{-F_1} \times e^{-F_2} \times e^{-F_3} \times \dots \times e^{-F_i} \times \dots \times e^{-F_n} \\ R_T = e^{-(F_1 + F_2 + F_3 + \dots + F_i + \dots + F_n)} \\ \text{Or } R_T = \exp \left(-\sum_{i=1}^n F_i \right) \end{aligned} \quad (2)$$

when $t_1 = t_2 = \dots t_n$, then Eq. (2) can be written as

$$R_T = \exp \left(-t \sum_{i=1}^n \lambda_i \right) \quad (2a)$$

Mean Time to Failure (MTTF)

In a constant-failure-rate series system, the mean time to failure m is the reciprocal of the system failure rate λ . The reliability of a component is its probability of survival. If a large number of components are put on test, then the reliability at any time t is equal to the ratio of the number of units still operating at that time (surviving). N_s , divided by the initial total number N_T , as given by Eq. 3.

$$R(t) = \frac{N_s}{N_T} \quad (3)$$

Since the number of units still operating N_s is equal to the total number N_T minus the number which have failed N_F , Eq. (3) may be rewritten as

$$R(t) = \frac{N_s}{N_T} = \frac{N_T - N_F}{N_T} = 1 - \frac{N_F}{N_T} \quad (4)$$

The total number of units N_T is constant while the number which have failed increases with time.

Hence, we can write:

$$\frac{dR}{dt} = \frac{d(1 - N_F/N_T)}{dt} = - \frac{1}{N_T} \frac{dN_F}{dt} \quad (5)$$

By rearranging the terms in Eq. (5) we obtain

$$\frac{1}{N_T} \frac{dN_F}{dt} = - \frac{dR}{dt} \quad (6)$$

In Eq. (6), dN_F/dt is the frequency at which failures occur when the total number of units N_T remains constant (no replacement of failed units.) When plotted on a graph as a function of time t , we obtain the time distribution of failures. If we divide dN_F/dt by the initial total number of units N_T , we obtain the distribution of failures, or failure frequency curve, *per component*. Such a unit failure distribution curve is called a *failure density function*, of simply $f(t)$. Substituting this term into Eq. (6), we obtain

$$f(t) = - \frac{dR}{dt} \quad (7)$$

Equation (7) applies to all possible failure density functions and not just to the case of a constant failure rate.

Now, the mean time to failure m , like any mean value, is the first moment about the origin

of the parameter being considered. In this case it is the average time at which failure occurs and can be found by operating all units to failure, summing the times to failure, and dividing by the number of units.

$$m = \frac{\sum_{i=1}^n \text{times to failure}}{n} = \frac{\sum_{i=1}^n t_i N_{Fi}}{N_T} \quad (8)$$

where $\sum N_{Fi} = N_T$. Equation (8) defines the average time to failure for one component. In the limit, as the number of units becomes infinitely large, the summation process becomes an integration process. Also, the distribution of failures per component N_{Fi}/N_T becomes defined by the failure density function $(d N_F/dt) N_T$, and m becomes the integral of the product of the density function and time:

$$m = \int_0^{\infty} t f(t) dt \quad (9)$$

Equation (9), like Eq. (7), is applicable to all possible density functions. Since $f(t) = -dR/dt$ [from Eq. (7)], Eq. (9) can be rewritten

$$m = \int_0^{\infty} t \left(-\frac{dR}{dt} \right) dt = \int_0^{\infty} -t dR \quad (10)$$

Integrating Eq. (10) by parts, we obtain

$$m = -[tR]_0^{\infty} + \int_0^{\infty} R dt \quad (11)$$

We can show that the first term on the right side of Eq. (11) is zero. (Details not included)

Therefore, Eq. (11) reduces to

$$m = \int_0^{\infty} R dt \quad (12)$$

This means that, in the general case, the mean time to failure can be obtained by integrating the reliability function over the time interval 0 to ∞ . Hence, for the specific case of a constant failure-rate component, we can write:

$$R = \exp\left(-\int_0^t \lambda_i dt\right) = \exp\left(-\lambda_i \int_0^t dt\right) = \exp(-\lambda_i t)$$

which, of course, we already know. From Eq. (12),

$$\begin{aligned} m &= \int_0^{\infty} R dt = \int_0^{\infty} e^{-\lambda_i t} dt = \int_0^{\infty} e^{\frac{-\lambda_i t - \lambda_i}{-\lambda_i}} dt \\ &= -\frac{1}{\lambda_i} [e^{-\infty} - e^0] = -\frac{1}{\lambda_i} (0 - 1) = \frac{1}{\lambda_i} \end{aligned} \quad (13)$$

In a series system, λ is merely the sum of the failure rates of the individual components, so that

$$m = \frac{1}{\lambda_1 + \lambda_2 + \dots + \lambda_n} = \frac{1}{\lambda} \quad (13a)$$

Parallel Systems

Two-Unit Parallel System - One Required

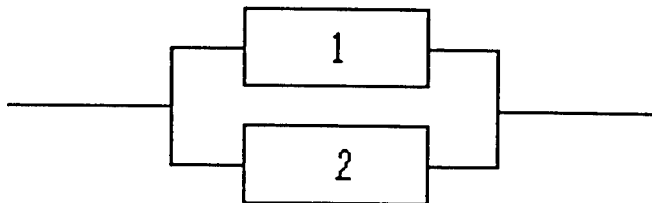


Fig. 2 Two-Unit Parallel System

In a simple parallel system consisting of two units in parallel with both operating but only one required, both must fail for the system to fail. The probability of the first unit failing is Q_1

which is equal to $1 - R_1 = 1 - e^{-F_1}$, and the probability of the second unit failing is $Q_2 = 1 - e^{-F_2}$.

The probability of both failing is

$$Q_1 \times Q_2 = (1 - e^{-F_1}) \times (1 - e^{-F_2})$$

This is better expressed as

$$= 1 - e^{-F_1} - e^{-F_2} + (e^{-F_1} \times e^{-F_2})$$

$$1 - e^{-F_1} - e^{-F_2} + e^{-(F_1 + F_2)} \quad (14)$$

The reliability is equal to one minus the probability of failure, or

$$R_T = e^{-F_1} + e^{-F_2} - e^{-(F_1 + F_2)} \quad (15)$$

In terms of the reliability of the individual units, the overall reliability can be expressed as

$$R_T = R_1 + R_2 - R_1 R_2 \quad (16)$$

An alternate form, derived directly from the fact that the system fails only when both units fail, is given by

$$R_T = 1 - (Q_1 \times Q_2) \quad (17)$$

When both units are the same, Eqs. (15), (16) and (17) simplify as follows:

$$R_T = 2e^{-F} - e^{-2F} \quad (15a)$$

$$R_T = 2R - R^2 \quad (16a)$$

$$R_T = 1 - Q^2 \quad (17a)$$

The mean time of failure (MTTF) for a two-unit parallel configuration is found by integrating the appropriate reliability function. Equation (15) is utilized in the integration, but in the form using λt rather than F .

$$m = \int_0^{\infty} R_T dt = \int_0^{\infty} (e^{-\lambda_1 t} + e^{-\lambda_2 t} - e^{-(\lambda_1 + \lambda_2)t}) dt \quad (18)$$

We can integrate each term independently. Since the failure rate for each individual component is constant, we have

$$\begin{aligned}
m &= \int_0^{\infty} e^{-\lambda_1 t} \times \frac{-\lambda_1}{-\lambda_1} dt + \int_0^{\infty} e^{-\lambda_2 t} \times \frac{-\lambda_2}{-\lambda_2} dt \\
&= \int_0^{\infty} e^{-(\lambda_1 + \lambda_2)t} \times \frac{-(\lambda_1 + \lambda_2)}{-(\lambda_1 + \lambda_2)} dt \\
&= \frac{1}{\lambda_1} \int_0^{\infty} e^{-\lambda_1 t} d(-\lambda_1 t) - \frac{1}{\lambda_2} \int_0^{\infty} e^{-\lambda_2 t} d(-\lambda_2 t) \\
&= \frac{1}{(\lambda_1 + \lambda_2)} \int_0^{\infty} e^{-(\lambda_1 + \lambda_2)t} d(-\lambda_1 + \lambda_2)t \\
&= \frac{1}{\lambda_1} \left[e^{-\lambda_1 t} \right]_0^{\infty} - \frac{1}{\lambda_2} \left[e^{-\lambda_2 t} \right]_0^{\infty} + \frac{1}{\lambda_1 + \lambda_2} \left[e^{-(\lambda_1 + \lambda_2)t} \right]_0^{\infty} \\
&= \frac{1}{\lambda_1} (0 - 1) - \frac{1}{\lambda_2} (0 - 1) + \frac{1}{\lambda_1 + \lambda_2} (0 - 1) \\
&= \frac{1}{\lambda_1} - \frac{1}{\lambda_2} - \frac{1}{\lambda_1 + \lambda_2}
\end{aligned} \tag{19}$$

In the case where $\lambda_1 = \lambda_2$, Eq. (19) reduces to

$$m = \frac{1}{\lambda} + \frac{1}{\lambda} - \frac{1}{2\lambda} = \frac{3}{2\lambda} \tag{19a}$$

Three-Unit Parallel System - One Required

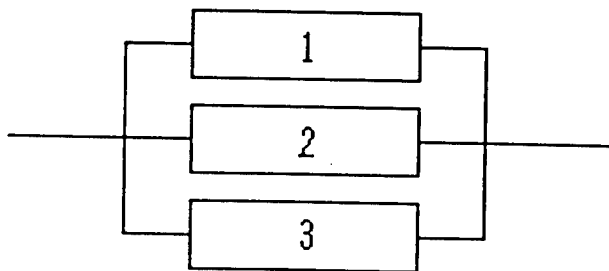


Fig. 3 Three-Unit Parallel System

When there are three units in parallel and only one is required as in Fig. 3, all three must fail for the system to fail, and the probability of system failure is

$$Q_1 \times Q_2 \times Q_3 = (1 - e^{-F_1}) \times (1 - e^{-F_2}) \times (1 - e^{-F_3})$$

This expands to

$$1 - e^{-F_1} - e^{-F_2} - e^{-F_3} + (e^{-F_1} \times e^{-F_2}) + (e^{-F_1} \times e^{-F_3}) + (e^{-F_2} \times e^{-F_3}) - (e^{-F_1} \times e^{-F_2} \times e^{-F_3})$$

The reliability is, of course, one minus the failure probability and can be written

$$R_T = e^{-F_1} + e^{-F_2} + e^{-F_3} - e^{-(F_1+F_2)} - e^{-(F_1+F_3)} - e^{-(F_2+F_3)} + e^{-(F_1+F_2+F_3)} \quad (20)$$

This, in turn, can be expressed as

$$R_T = R_1 + R_2 + R_3 - (R_1 \times R_2) - (R_1 \times R_3) - (R_2 \times R_3) + (R_1 \times R_2 \times R_3) \quad (21)$$

In terms of failure probabilities,

$$R_T = 1 - (Q_1 \times Q_2 \times Q_3) \quad (22)$$

Equations (20), (21), and (22) for three units in parallel with only one required reduce to

Eqs. (20a), (21a) and (22a), when the three units are alike.

$$R_T = 3e^{-F} - 3e^{-2F} + e^{-3F} \quad (20a)$$

$$R_T = 3R - 3R^2 + R^3 \quad (21a)$$

$$R_T = 1 - Q^3 \quad (22a)$$

For a three-unit parallel system where only one is required to operate, the MTTF is computed as follows. In the general case when the failure rates are different, Eq. (20) is used in the integration with λt used rather than F .

$$m = \frac{1}{\lambda_1} + \frac{1}{\lambda_2} + \frac{1}{\lambda_3} - \frac{1}{\lambda_1 + \lambda_2} - \frac{1}{\lambda_1 + \lambda_3} - \frac{1}{\lambda_2 + \lambda_3} + \frac{1}{\lambda_1 + \lambda_2 + \lambda_3} \quad (23)$$

When the failure rates are the same, Eq. (20a) is used.

$$m = \frac{3}{\lambda} - \frac{3}{2\lambda} + \frac{1}{3\lambda} = \frac{18}{6\lambda} - \frac{9}{6\lambda} + \frac{2}{6\lambda} = \frac{11}{6\lambda} \quad (23a)$$

Multiple-Unit Parallel System - One Required

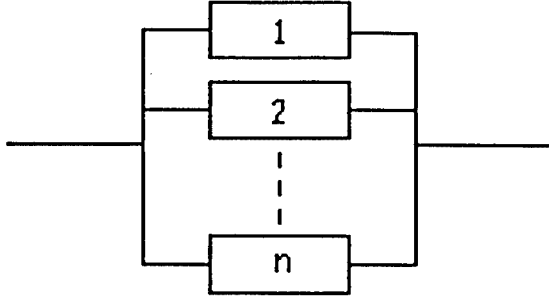


Fig. 4 Multiple-Unit Parallel System

In the large majority of cases, when there are more than three units in parallel, all units are the same. Consequently the following derivations consider only cases of identical units.

The equations for the reliability and mean time to failure for additional redundancy when only one operable unit is required can be found by expanding the derivations for Eqs. (20) through (23a) to cover the additional units. In the equations, n is the total number of units and $n-1$ failures are permitted. The three equivalent reliability expressions are

$$R_T = ne^{-F} - \frac{n(n-1)}{2} e^{-2F} + \frac{n(n-1)(n-2)}{2 \times 3} e^{-3F} - \frac{n(n-1)(n-2)(n-3)}{2 \times 3 \times 4} e^{-4F} + \dots (+) e^{-nF} \quad (24)$$

$$R_T = nR - \frac{n(n-1)}{2} R^2 + \frac{n(n-1)(n-2)}{3!} R^3 - \frac{n(n-1)(n-2)(n-3)}{4!} R^4 + \dots (+) R^n \quad (25)$$

$$R_T = 1 - Q^n \quad (26)$$

The mean time of failure, derived from the complete expansion of Eq. (24) after integrating, combining terms, and simplifying is found to be

$$m = \frac{1}{\lambda} + \frac{1}{2\lambda} + \frac{1}{3\lambda} + \frac{1}{4\lambda} + \dots + \frac{1}{n\lambda} \quad (27)$$

It can be seen from the preceding equations [and from Eq. (26) in particular] that as more units are added in parallel, the total reliability rapidly approaches unity, differing only by Q^n . However, as shown by Eq. (27), m does *not* increase correspondingly and each additional parallel redundant unit contributes less to the system MTTF than did its predecessor.

Combination Systems

The organization structure of most systems can be described as a combination of series and parallel systems. The logic diagram of such a system is shown in Fig. 5.

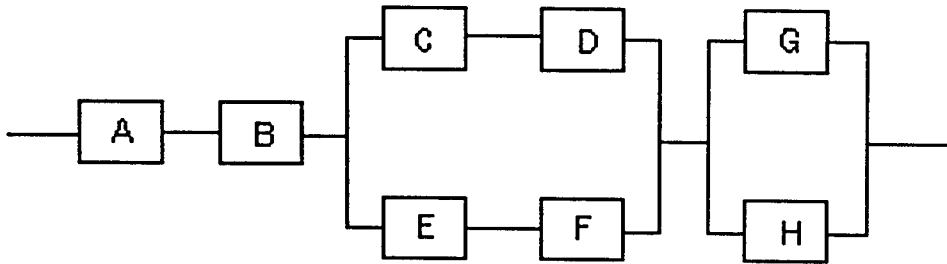


Fig. 5 Logic Diagram of a Combination System

The precision of reliability prediction for these types of combination system can be determined by the application of Limiting Value technique (Method of Bounds). The model determines the limiting values (upper and lower bounds) of reliability of the system and obtains a single system reliability prediction value, utilizing the upper and lower bounds. The model is described below:

Calculation of the Upper Bound

The calculation of the upper bound considers only those elements failures which individually would cause mission failure. Hence, only those blocks in the logic diagram which are in series are considered. Generally, this is sufficient to provide a satisfactory estimate. However, if the reliabilities associated with individual blocks are not very high, it may be necessary to consider system failures resulting from multiple component failures, i.e. to consider the parallel blocks. When only the series elements are considered, the system reliability upper bound R_{upper} is the product of the reliabilities of elements A and B.

$$R_{upper} = R_A \times R_B$$

Assuming an exponential model (constant-failure-rate system), the reliabilities of the components can be expressed as:

$$R_A = e^{-F_A} \text{ and } R_B = e^{-F_B}$$

where F_A and F_B are the mission failure rates of elements A and B respectively. Therefore,

$$R_{upper} = e^{-F_A} \times e^{-F_B} = e^{-(F_A + F_B)}$$

In general, the upper bound can be calculated using the following equation

$$R_{upper} = \exp \left(- \sum_{i=1}^m F_i \right) \quad (28)$$

where F_i is the mission failure rate of series element i , and m is the number of series elements.

It is seldom necessary to consider parallel elements. However, when the reliabilities of some elements are relatively low (e.g., less than 0.99), and these are functionally in parallel, the system reliability upper bound, if only series elements are considered, may be overly optimistic. It can be lowered by considering multiple failures. In Figure 5, failure of any of the following pairs

of elements results in mission failure: C and E, C and F, D and E, D and F, or G and H. If these components are not highly reliable, there may be more than a negligible chance of both elements in a pair failing. In that case, two elements at a time should be considered.

When system failure results from failure of two parallel elements, all series elements must be good (otherwise the system would have failed due to the failure of a series element). The series elements must therefore be included in these calculations. Hence, system failure probability resulting from failure, for example, of elements C and E in the preceding logic diagram is defined as $R_A \times R_B \times Q_C \times Q_E$, where $Q = 1 - R$. Similar expressions apply to the other pairs of parallel elements enumerated above.

$$\begin{array}{ll} R_A \times R_B \times Q_C \times Q_F & R_A \times R_B \times Q_D \times Q_E \\ R_A \times R_B \times Q_D \times Q_F & R_A \times R_B \times Q_G \times Q_H \end{array}$$

The probability of system failure resulting from the failure of two non-series elements is found by adding these terms.

$$P = (R_A \times R_B \times Q_C \times Q_E) + (R_A \times R_B \times Q_C \times Q_F) + \dots \\ + (R_A \times R_B \times Q_G \times Q_H)$$

This can be simplified to

$$P = R_A \times R_B \times [(Q_C \times Q_E) + (Q_C \times Q_F) + \dots + (Q_G \times Q_H)]$$

Since $R_A \times R_B = e^{-F_i}$, from Eq. (28), we can write this as

$$\left[\exp \left(- \sum_{i=1}^m F_i \right) \right] \left[\sum_{(k, k')=1}^x (Q_k \times Q_{k'}) \right] \quad (29)$$

where m is the number of series elements, Q_k and $Q_{k'}$ are the failure *probabilities* of pairs of parallel elements which together cause system failure, and x is the number of such pairs. This

value is subtracted from Eq. (28) to obtain the system reliability upper bound when two failures are considered:

$$R_{upper} = \left[\exp \left(- \sum_{i=1}^m F_i \right) \right] - \left[\exp \left(- \sum_{i=1}^m F_i \right) \right] \left[\sum_{(k,k')=1}^x (Q_k \times Q_{k'}) \right]$$

$$R_{upper} = \left[\exp \left(- \sum_{i=1}^m F_i \right) \right] \left[1 - \sum_{(k,k')=1}^x (Q_k \times Q_{k'}) \right] \quad (30)$$

Calculation of the Lower Bound

The lower bound is found by adding probabilities of success cases. In redundant systems, there are usually many cases where one or more element failures can occur without causing system failure. Therefore, it is always necessary when calculating the lower reliability bound of a redundant system to consider all cases with one failure, and frequently cases with two or even three failures.

The first calculation considers the case of no failures, that is, where all elements are good. This is simply the product of the reliabilities of all elements,

$$R_{lower} = \prod_{j=1}^n R_j$$

where R_j is the reliability of any element, *series or parallel*, and n is the total number of elements. In the previous logic diagram of Fig. 5, the first calculation of the lower bound is equal to $R_A \times R_B \times R_C \times \dots \times R_H$.

$$= e^{-F_A} \times e^{-F_B} \times e^{-F_C} \times \dots \times e^{-F_H} = \exp \left(- \sum_{A}^H F \right)$$

The general case for zero failures is expressed as follows:

$$R_{lower} = \exp \left(- \sum_{j=1}^n F_j \right) \quad (31)$$

Thus, a failure of any one element, C, D, E, F, G, or H is still a system success. These cases are listed here for clarity, with R denoting success and Q denoting failure.

$$\begin{aligned}
& (R_A \times R_B \times Q_C \times R_D \times R_E \times R_F \times R_G \times R_H) \\
& + (R_A \times R_B \times R_C \times Q_D \times R_E \times R_F \times R_G \times R_H) \\
& + \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \\
& \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \\
& \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \\
& + (R_A \times R_B \times R_C \times R_D \times R_E \times R_F \times R_G \times Q_H)
\end{aligned}$$

These expressions can be simplified by multiplying by appropriate terms equal to unity and then simplifying, as follows:

$$\left(R_A \times R_B \times Q_C \times \frac{R_C}{P} \times R_D \times R_E \times R_F \times R_G \times R_H \right)$$

$$\begin{aligned}
&= \left(R_A \times R_B \times Q_C \times \frac{R_C}{R_C} \times R_D \times R_E \times R_F \times R_G \times R_H \right) \\
&+ \left(R_A \times R_B \times R_C \times Q_D \times \frac{R_D}{R_D} \times R_E \times R_F \times R_G \times R_H \right) \\
&+ \dots \\
&+ \left(R_A \times R_B \times R_C \times R_D \times R_E \times R_F \times R_G \times Q_H \times \frac{R_H}{R_H} \right) \\
&= \left(R_A \times R_B \times R_C \times R_D \times R_E \times R_F \times R_G \times R_H \times \frac{Q_C}{R_C} \right) \\
&+ \left(R_A \times R_B \times R_C \times R_D \times R_E \times R_F \times R_G \times R_H \times \frac{Q_D}{R_D} \right) \\
&+ \dots \\
&+ \left(R_A \times R_B \times R_C \times R_D \times R_E \times R_F \times R_G \times R_H \times \frac{Q_H}{R_H} \right) \\
&= (R_A \times R_B \times R_C \times R_D \times R_E \times R_F \times R_G \times R_H) \\
&\times \left(\frac{Q_C}{R_C} + \frac{Q_D}{R_D} + \dots + \frac{Q_H}{R_H} \right)
\end{aligned}$$

The first term is simply the probability of all elements being good, as expressed in Eq. (31). The second term is the sum of the ratios of failure probability to success probability of all non-series elements. The product of these two terms can be written as

$$= \left[\exp \left(- \sum_{j=1}^n F_j \right) \right] \left[\sum_{k=1}^p \left(\frac{Q_k}{R_k} \right) \right] \quad (32)$$

where Q_k is the probability of failure of a non-series element, R_k is the reliability of the non-series element, and p is the number of non-series elements of the system. The equation for cases of two non-series element failures which together *do not* cause system failure is similarly derived:

$$= \left[\exp \left(- \sum_{j=1}^n F_j \right) \right] \left[\sum_{(k,l)=1}^{p'} \left(\frac{Q_k}{R_k} \times \frac{Q_l}{R_l} \right) \right] \quad (33)$$

where k and l are pairs of non-series elements which, failing together, do not cause system failure and p' is the number of such pairs. Equations (31), (32), and (33) are then added to give the lower-bound probability of success considering no element failures, one element failure, and two element failures.

$$\begin{aligned} R_{\text{lower}} &= \exp \left(- \sum_{j=1}^n F_j \right) + \left[\exp \left(- \sum_{j=1}^n F_j \right) \right] \left[\sum_{k=1}^p \left(\frac{Q_k}{R_k} \right) \right] \\ &+ \left[\exp \left(- \sum_{j=1}^n F_j \right) \right] \left[\sum_{(k,l)=1}^{p'} \left(\frac{Q_k}{R_k} \times \frac{Q_l}{R_l} \right) \right] \\ &= \left[\exp \left(- \sum_{j=1}^n F_j \right) \right] \left[1 + \sum_{k=1}^p \left(\frac{Q_k}{R_k} \right) + \sum_{(k,l)=1}^{p'} \left(\frac{Q_k}{R_k} \times \frac{Q_l}{R_l} \right) \right] \end{aligned} \quad (34)$$

Final Combined Prediction

The last step is the combining of the two bounds to obtain a single system reliability prediction value. The easiest method would be to take the simple arithmetic average of the two bounds. However, experience has shown that this results in an overly pessimistic value. It has been found empirically that the true value of the system failure probability can be more closely approximated by taking the square root of the product of the unreliabilities associated with the upper and lower reliability bounds. This is then subtracted from unity to obtain the single prediction value.

$$R_{\text{system}} = 1 - \sqrt{(1 - R_{\text{upper}})(1 - R_{\text{lower}})} \quad (35)$$

It is important that the calculations of the two bounds be stopped at the same point. That is, if only cases of one failure are considered for the upper bound, then cases of only zero and one failure should be considered for the lower bound; if two failures are considered for the upper bound, then consider cases of two failures for the lower bound, etc. Inaccuracies will result in Eq. (35), if the calculations are not carried to the same point.

Results and Discussions

The computational data used in this investigation for the models described in the previous section were made available by the Reliability Analysis Center (RAC), Rome, N.Y. The data used are presented in the appendix. Two types of capacitors were used and they are listed below.

Table 1

| Capacitor Component Reliability Results | | | | | |
|---|-------------------------------|---------------|-------------------------|-------------|-------------|
| Part Number | Part Description | Quality Level | Application Environment | Data Source | Reliability |
| 1 | Fixed, electrolytic, tantalum | Commercial | GB | 13567-021 | 0.9947 |
| 2 | Fixed, electrolytic, tantalum | Military | GF | 14851-000 | 0.9938 |

It may be noted that because of the limited nature of the available field data, the component reliability results shown in the table should be treated with caution. A degree of uncertainty exists in the data used for the computation. The reliability results for various forms of series and parallel systems are shown in Table 2. The results indicate that, as expected, the reliability of the parallel system rapidly approaches unity, when more components are added.

The reliability and bounds for the combination (series and parallel) system are presented in Table 3. It is also pictorially shown in Fig. 6.

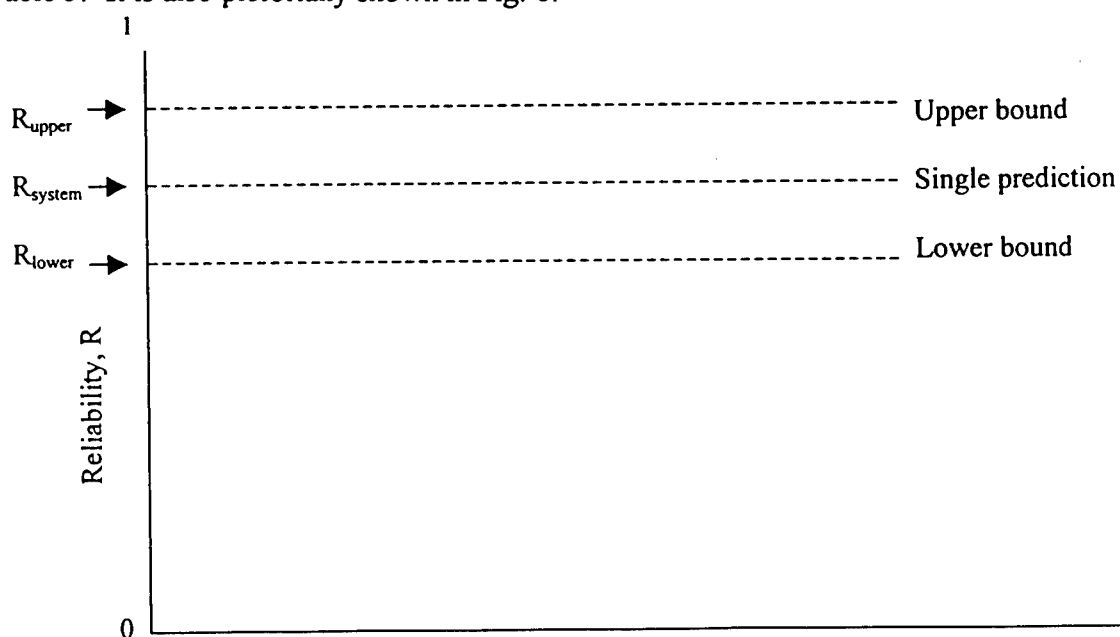


Fig. 6 Upper and lower bounds for the combination system

Here the equations of the mathematical models are precise; however, the type of results shown is always questionable because of the inherent uncertainty of the data used. This well known problem in reliability engineering was addressed by Zadeh (Ref. 5, 6), Dempster (Ref. 8) and Shafer (Ref. 9). Their work in general is known as possibility theories. An attempt has been made to apply this new theory in the current investigation. The details of this endeavor are described in the following section.

Table 2

| Reliabilities for Series and Parallel Systems | | | | |
|---|-------------------------|-------------------|-------------|-------------|
| Figure Numbers | System | No. of Components | Part Number | Reliability |
| 1 | Series | 7 | 1 | 0.9635 |
| 1 | Series | 7 | 2 | 0.9574 |
| 2 | Parallel - one required | 2 | 1 | 0.99997 |
| 2 | Parallel - one required | 2 | 2 | 0.99996 |
| 3 | Parallel - one required | 3 | 1 | 0.99999 |
| 3 | Parallel - one required | 3 | 2 | 0.99999 |
| 4 | Parallel - one required | 7 | 1 | 0.99999 |
| 4 | Parallel - one required | 7 | 2 | 0.99999 |

Table 3

| Reliabilities and Bounds for Combination (Series & Parallel) System | | | |
|--|-------------------------|-------------------------|--------------------------|
| Number of components = 8 (Fig. 5) | | | |
| Case I: All components of the system are good. | | | |
| Part Number | Upper Bound R_{upper} | Lower Bound R_{lower} | Reliability R_{system} |
| 1 | 0.9894 | 0.9584 | 0.9790 |
| 2 | 0.9876 | 0.9515 | 0.9755 |
| Case II: One non-series component fails and all other components are good. | | | |
| Part Number | R_{upper} | R_{lower} | R_{system} |
| 1 | 0.9894 | 0.9890 | 0.9892 |
| 2 | 0.9876 | 0.9871 | 0.9874 |

The Possibility Theory (Task 3)

In the mathematical models described in the previous section for the determination of reliabilities and their precision, it should be noted that although the equations are precise, the statistical data used in the computation seldom are. A consequence of this is the inherent uncertainty in all reliability results. The uncertainty is mainly due to the fact that failures being relatively rare events (typically only a few per million hours of operation), collecting enough data on which to base a statistical "probability of failure" is a costly and difficult undertaking, and the relevance of the data to any particular system, as well as its validity, is often questionable. Further extrapolating these failure probabilities through statistical methods to calculate a system level reliability only increases the uncertainty.

This led to the development of the possibility theories, commonly known as "Fuzzy Sets" theory (FST) and Evidence Theory (ET). The Fuzzy Set Theory was originally presented by

Zadeh (Ref. 5, 6) and provides the basis of a possibilities approach to system reliability evaluation based on the premise that a small probability does not always mean a low possibility of an event, whereas a low possibility would necessarily imply a low probability (Ref. 7). FST may be thought of as a generalized form of the conventional Boolean Set Theory. The difference is that an object can have a Fuzzy Set Membership, $\mu(x)$, anywhere in the continuous range of 0 to 1, but for Boolean Sets, membership is restricted to values exactly equal to 0 or 1.

The Evidence Theory (ET) originated from the work of Dempster (Ref. 8) and was developed by Shafer (Ref. 9). It is a new tool for representing situations in which various kinds of ignorance exist in our knowledge or information about a system. It is a reasoning approach for testing multiple hypothesis on the basis of evidence. The only assumption made about the evidence is that the sum of the values supporting all conclusions plus the unknown equals 1. Evidence from multiple sources is combined by a geometric procedure that generates the mass of evidence supporting each possible conclusion.

Fuzzy Numbers and the Relevant Arithmetic

The concept of uncertain or fuzzy numbers may be considered as an extension of the concept of the Interval of confidence. This extension is based on a natural and very simple idea. Instead of considering the interval of confidence at one unique level, it is considered at several levels and more generally at all levels from 0 to 1 (Ref. 10). Here we consider the maximum of presumption to be at level 1 and the minimum of presumption to be at level 0. It should be noted that the fuzzy numbers are not random numbers (more commonly known as Random Variables). Uncertainty and randomness are two very different and important concepts. They can be used together but should not be confused.

Fuzzy numbers are numerical approximations such as “about 5”. For simplicity fuzzy numbers are often represented with triangular membership functions, $\mu(x)$, as shown in Fig. 7.

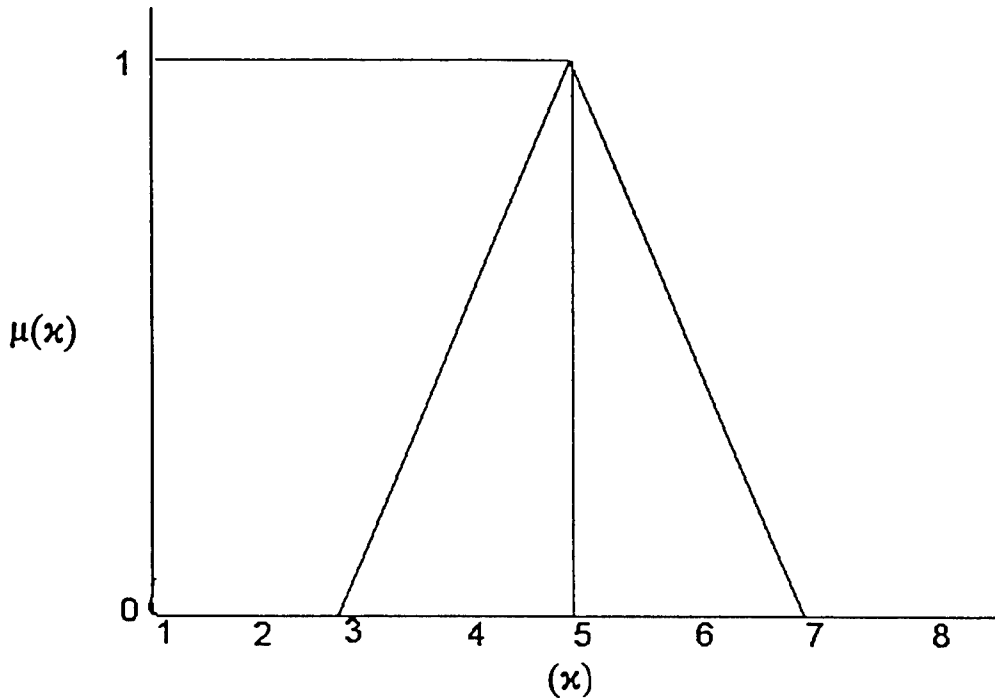


Fig. 7

The width of the membership function shows the range of possible values. The number in Fig. 7 can be interpreted as ranging from 3 to 7 with the values near 5 being most similar to 5 or better satisfying the membership property: “close to 5”, than those farther away (Ref. 11, 12). At this stage, probably it will be desirable to relate the concept of the Interval of Confidence to the Level of Presumption with a simple example.

Suppose a certain job is to be completed between two dates, say June 10 and June 30. This is an interval of confidence. On the other hand, let us assume that this same job is to be completed on June 20, a possible date. The interval of confidence in the first case is [June 10, June 30] while in the second case it is [June 20, June 20]. If we wish, we may assign two levels

of confidence to these two situations, 0 for [June 10, June 30] and 1 for [June 20, June 20].

These two levels of confidence are in fact Levels of Presumption, and we can represent them by [0,1]. Of course, there is no reason why we should limit ourselves to only the two values 0 and 1.

We could have selected a much larger set of values such as:

$$\forall \alpha_1, \alpha_2 \in [0,1]:$$

$$(\alpha_1 < \alpha_2) \rightarrow ([a_1^{(\alpha_2)}, a_2^{(\alpha_2)}] \subset [a_1^{(\alpha_1)}, a_2^{(\alpha_1)}]).$$

This means that if α increases, the interval of confidence never increases. Fig. 8 illustrates graphically the mathematical relation. It may be noted that the different values of $\alpha_1, \alpha_2, \dots$ etc. are commonly called α - cuts (Ref. 10).

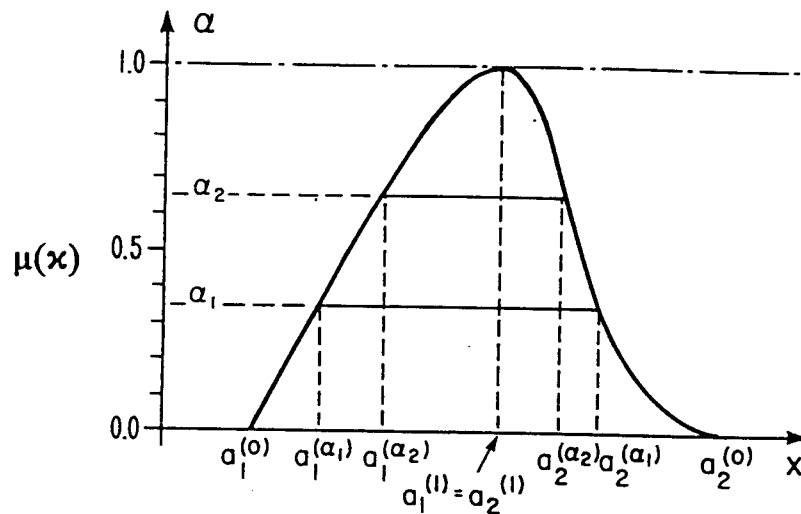


Fig. 8 Definition of fuzzy numbers

The Basic Operations on Fuzzy Numbers [Ref. 11, 12]

$$A_\alpha + B_\alpha = [a_1^\alpha, a_2^\alpha] + [b_1^\alpha, b_2^\alpha]$$

$$= [a_1^\alpha + b_1^\alpha, a_2^\alpha + b_2^\alpha] = [c_1^\alpha, c_2^\alpha] \quad (1)$$

$$A_\alpha - B_\alpha = [a_1^\alpha, a_2^\alpha] - [b_1^\alpha, b_2^\alpha]$$

$$= [a_1^\alpha - b_2^\alpha, a_2^\alpha - b_1^\alpha] = [c_1^\alpha, c_2^\alpha] \quad (2)$$

$$\begin{aligned}
A_\alpha \times B_\alpha &= [a_1^\alpha, a_2^\alpha] \times [b_1^\alpha, b_2^\alpha] \\
&= [\min(a_1^\alpha b_1^\alpha, a_1^\alpha b_2^\alpha, a_2^\alpha b_1^\alpha, a_2^\alpha b_2^\alpha), \\
&\quad \max(a_1^\alpha b_1^\alpha, a_1^\alpha b_2^\alpha, a_2^\alpha b_1^\alpha, a_2^\alpha b_2^\alpha)] \\
&= [c_1^\alpha, c_2^\alpha]
\end{aligned} \tag{3a}$$

$$\begin{aligned}
A_\alpha / B_\alpha &= [a_1^\alpha, a_2^\alpha] / [b_1^\alpha, b_2^\alpha] \\
&= [\min(a_1^\alpha / b_1^\alpha, a_1^\alpha / b_2^\alpha, a_2^\alpha / b_1^\alpha, a_2^\alpha / b_2^\alpha), \\
&\quad \max(a_1^\alpha / b_1^\alpha, a_1^\alpha / b_2^\alpha, a_2^\alpha / b_1^\alpha, a_2^\alpha / b_2^\alpha)] \\
&= [c_1^\alpha, c_2^\alpha]
\end{aligned} \tag{4a}$$

$$\begin{aligned}
1/B_\alpha &= 1/[b_1^\alpha, b_2^\alpha] \\
&= [\min(1/b_2^\alpha, 1/b_1^\alpha), \max(1/b_2^\alpha, 1/b_1^\alpha)] \\
&= [c_1^\alpha, c_2^\alpha]
\end{aligned}$$

$$\begin{aligned}
A_\alpha \pm k &= [a_1^\alpha, a_2^\alpha] + k \\
&= [a_1^\alpha \pm k, a_2^\alpha \pm k] = [c_1^\alpha, c_2^\alpha]
\end{aligned} \tag{5a}$$

$$A_\alpha \times k = [a_1^\alpha, a_2^\alpha] \times k \tag{6}$$

$$\begin{aligned}
&= [\min(ka_1^\alpha, ka_2^\alpha), \max(ka_1^\alpha, ka_2^\alpha)] \\
&= [c_1^\alpha, c_2^\alpha]
\end{aligned} \tag{7}$$

For $a_1^\alpha, a_2^\alpha, b_1^\alpha, b_2^\alpha, k \geq 0$ we get the following simplifications:

$$A_\alpha \times B_\alpha = [a_1^\alpha b_1^\alpha, a_2^\alpha b_2^\alpha] \tag{3b}$$

$$A_\alpha / B_\alpha = [a_1^\alpha / b_2^\alpha, a_2^\alpha / b_1^\alpha] \tag{4b}$$

$$1/B_\alpha = [1/b_2^\alpha, 1/b_1^\alpha] \tag{5b}$$

$$A_\alpha \times k = [ka_1^\alpha, ka_2^\alpha] \tag{7b}$$

The operations in (4) and (5) are undefined if the interval B_α contains 0. In the limit, as

b_1^α or b_2^α goes to 0 and the resulting interval becomes infinite.

Analysis of Series System

The reliability of a series system with multiple components was discussed in detail in the previous sections (Fig. 1). In this type of system all components must be operational and for a system with n component, the reliability is given by:

$$R_{sys} = R_1 \times R_2 \times \dots \times R_n \tag{8}$$

In terms of the failure probabilities, it can be expressed as

$$Q_{sys} = 1 - [(1-Q_1)(1-Q_2) \dots (1+Q_n)] \tag{9}$$

In this analysis, we have used two fixed, electrolytic, tantalum capacitors, identified as Part 1 (Data Source: 13567-021-commercial) and Part 2 (Data Source: 14851-000-military) (see Table 1).

For a two-component series systems equations 8 and 9 can be written as:

$$R_{sys} = R_1 \times R_2 \quad (8a)$$

$$Q_{sys} = 1 - [(1-Q_1)(1-Q_2)] \quad (9a)$$

Although the equations are precise, the statistical data used (see Appendix) for computation are not. Hence, applying fuzzy logic, we can treat the reliability and the probability of failure as fuzzy numbers. Hence,

$$\check{R}_{sys} = \check{R}_1 \times \check{R}_2 \quad (10)$$

$$\tilde{Q}_{sys} = 1 - [(\tilde{1}-\tilde{Q}_1)(\tilde{1}-\tilde{Q}_2)] \quad (11)$$

where the intervals are as follows:

$$\begin{array}{ll} \check{R}_1 & [r_{11}, r_{12}] \\ \check{R}_2 & [r_{21}, r_{22}] \\ \check{R}_{sys} & [r_{s1}, r_{s2}] \\ \tilde{Q}_{sys} & [q_{s1}, q_{s2}] \end{array}$$

$$\begin{array}{ll} \text{where } r_{s1} &= r_{11} \times r_{21} \\ r_{s2} &= r_{12} \times r_{22} \\ q_{s1} &= (1 - r_{s2}) \\ q_{s2} &= (1 - r_{s1}) \end{array}$$

The results of this analysis are shown in Table 4 and Fig. 9.

Table 4

| Two Component Series System | | | | | | | | | | | |
|--|----------|----------|---------------|----------|----------|--|----------|----------|-------------------|----------|----------|
| Fuzzy Reliability = \check{R}_1 = Part 1 (see Table 1) | | | | | | Fuzzy Reliability = \check{R}_2 = Part 2 | | | | | |
| \check{R}_1 | r_{11} | r_{12} | \check{R}_2 | r_{21} | r_{22} | \check{R}_{sys} | r_{s1} | r_{s2} | \tilde{Q}_{sys} | q_{s1} | q_{s2} |
| 0.9947 | 0.99 | 0.995 | 0.9938 | 0.99 | 0.995 | 0.9885 | 0.9801 | 0.99 | 0.0115 | 0.01 | 0.0199 |

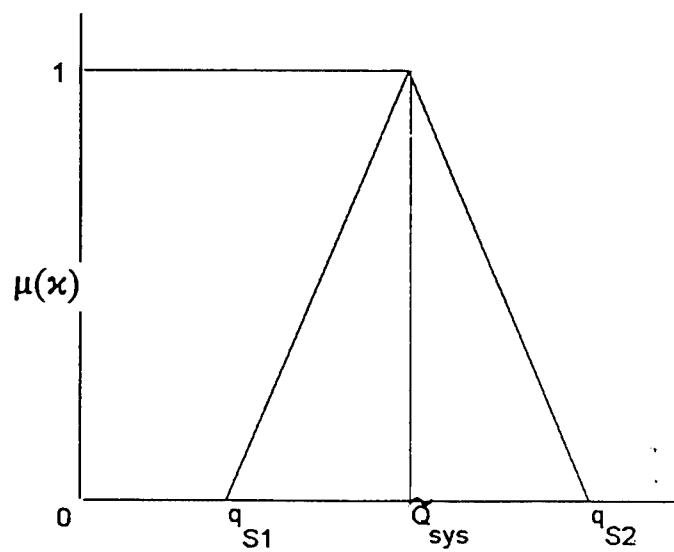
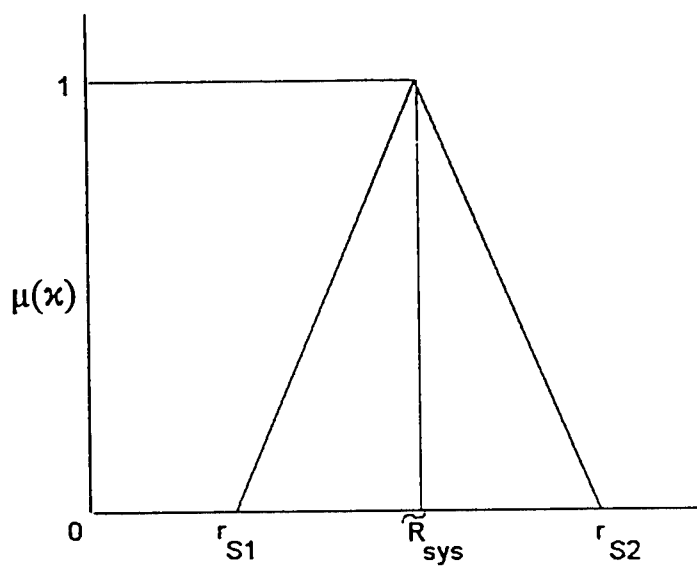
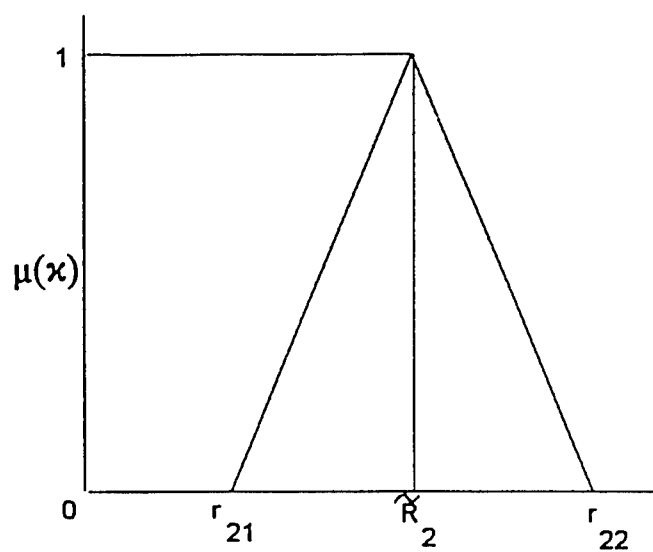
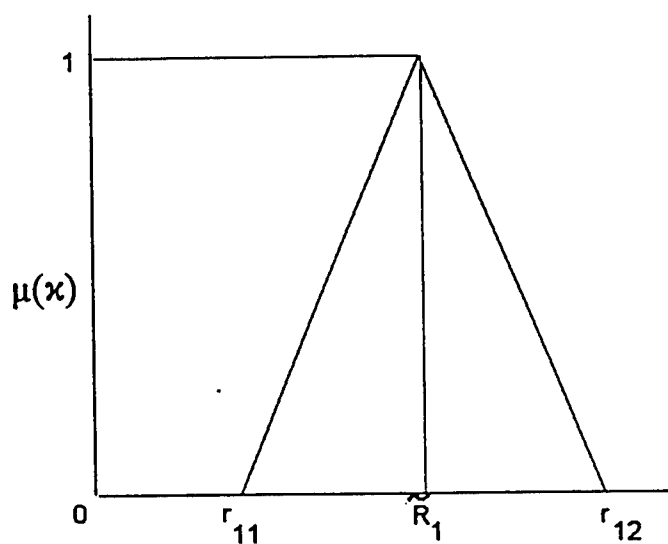


Fig. 9 Reliability and Failure Probability of a Two-component Series System

Analysis of Parallel System

The reliability of a multi-unit parallel system was also discussed in detail in previous sections (Fig. 4). Here the system works as long as at least one component is operational. The system failure probability can be written similarly as:

$$Q_{sys} = Q_1 \times Q_2 \times \dots \times Q_n \quad (12)$$

This equation can be rewritten for a two component - one required (Fig. 2) system as:

$$Q_{sys} = Q_1 \times Q_2 \quad (13)$$

Again, although the equation is precise, the statistical data used for the computation are not.

Hence, considering the failure probabilities as fuzzy numbers, we can write

$$\begin{aligned} Q_1 &\equiv \tilde{Q}_1 \\ Q_2 &\equiv \tilde{Q}_2 \\ Q_{sys} &\equiv \tilde{Q}_{sys} \end{aligned} \quad (14)$$

and hence

$$\tilde{Q}_{sys} = \tilde{Q}_1 \times \tilde{Q}_2 \quad (15)$$

where the intervals are:

$$\begin{aligned} \tilde{Q}_1 & [q_{11}, q_{12}] \\ \tilde{Q}_2 & [q_{21}, q_{22}] \\ \tilde{Q}_{sys} & [q_{s1}, q_{s2}] \end{aligned} \quad (16)$$

The system interval $[q_{s1}, q_{s2}]$ is given by

$$\begin{aligned} q_{s1} &= q_{11}, q_{21} \\ q_{s2} &= q_{12}, q_{22} \end{aligned} \quad (17)$$

The results of this analysis are shown in Table 5 and Fig. 10.

Table 5

| Two Component Parallel System | | | | | | | | |
|--|----------|----------|---------------|--|----------|------------------------|-------------------------|------------------------|
| Fuzzy Probability = \tilde{Q}_1 = Part 1 (see Table 1) | | | | Fuzzy Probability = \tilde{Q}_2 = Part 2 | | | | |
| \tilde{Q}_1 | q_{11} | q_{12} | \tilde{Q}_2 | q_{21} | q_{22} | \tilde{Q}_{sys} | q_{s1} | q_{s2} |
| 0.0053 | 0.00265 | 0.00795 | 0.0062 | 0.0031 | 0.0093 | 3.286×10^{-5} | 0.8215×10^{-5} | 7.394×10^{-5} |

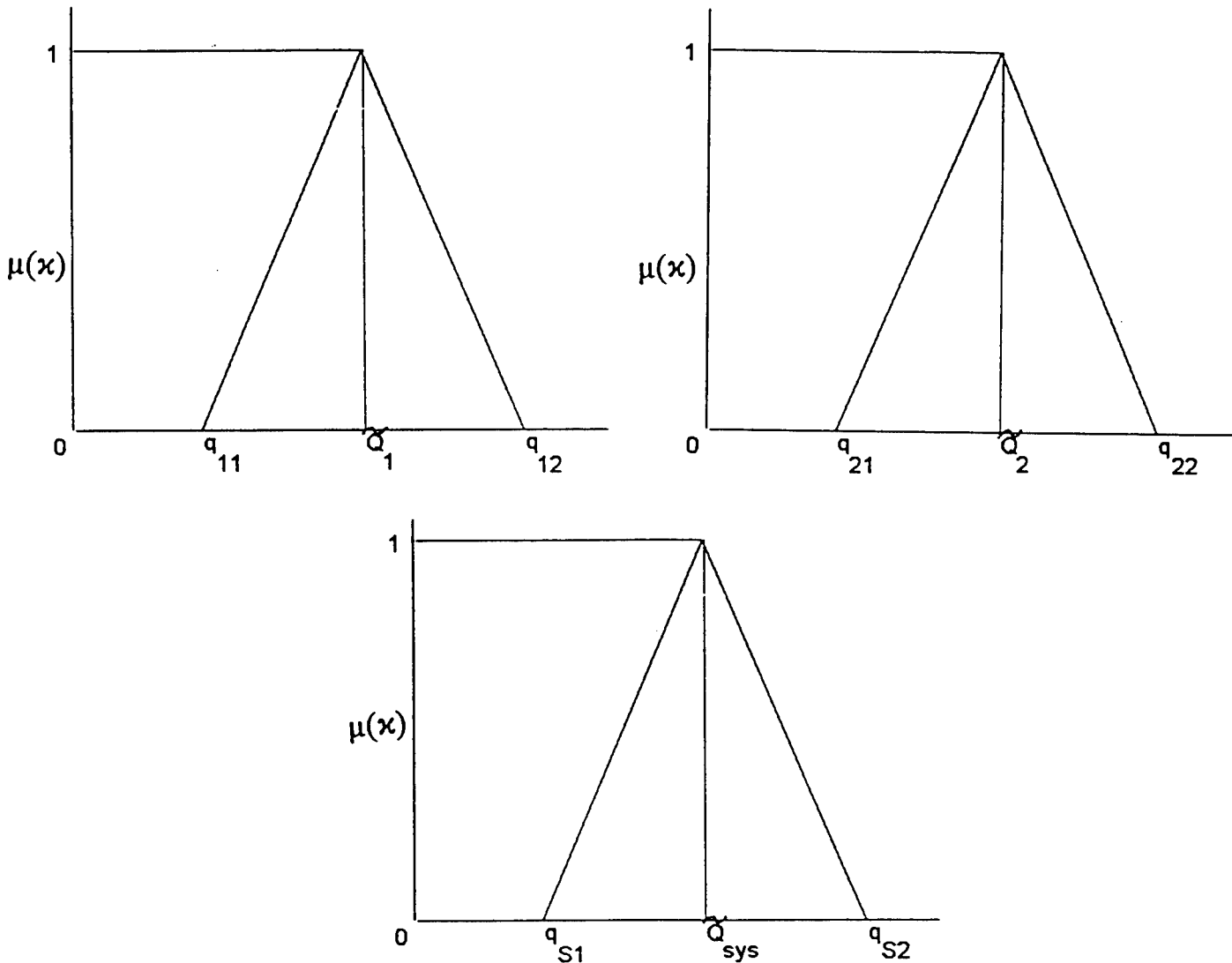


Fig. 10 Failure Probability of a Two Component Parallel System

It should be noted that though the results of this analysis are presented in the triangular membership function, $\mu(x)$, in practice, the membership functions could be a smooth curve similar to that shown in Fig. 8. Also, in the present work, the concept of a fuzzy number is presented using the couple: level of presumption and interval of confidence. This is not, however, the classical way of introducing a fuzzy number. The classical technique was not considered in the present work; however, the details of this technique are readily available in Ref. 10.

Conclusion

A simulation model for determining the precision of reliability was developed for a combination (series and parallel) system using the method of bounds. The model was validated with available field data. It is well known that field data inherently carry a certain amount of uncertainty. The model was developed further using the fuzzy number techniques to account for the uncertainty of the field data.

Acknowledgments

The author would like to extend his sincere thanks to Mr. Joseph A. Caroli for his support towards this project as the focal point at the Air Force Research Laboratory, Rome, NY. Thanks are also due to Mr. William Denson of Reliability Analysis Center (RAC), Rome, NY, for supplying the field data for this project and Ms. Eileen D'Agostino for the care she took in typing this report.

REFERENCES

1. Denson, William and Keene, Samuel, "New System Reliability Assessment Methods." RAC Project #A06830, January 1996.
2. Denson, William and Keene, Samuel, "A New System Reliability Assessment Method." Reliability Review, Vol. 15, p. 16-20, December 1995.
3. Das, D. K. "Techniques for Determining the Precision of Reliability Predictions and Assessments". SFRP Final Report, August 1996.
4. Amstadter, B. L., "Reliability Mathematics: Fundamentals, Practices, Procedures." McGraw Hill Book Company.
5. Zadeh, L.A., "Fuzzy Sets." Information and Control, Vol. 8, pg. 338-353, 1965.
6. Zadeh, L. A., "Fuzzy Sets as a Basis for a Theory of Possibility." Fuzzy Sets and Systems, Vol. 1, No. 1, p. 3-28, 1978.
7. Misra, K. B. (editor), "New Trends in System Reliability Evaluation." Elsevier Science Publisher, 1993.
8. Dempster, A. P., "Upper and Lower Probabilities Induced by a Multi-Valued Mapping." Ann. Math. Statist., Vol. 38, p. 325-339, 1967.
9. Shafer, G., "A Mathematical Theory of Evidence." Princeton University Press, Princeton, New Jersey, 1976.
10. Kaufmann, A. and Gupta, M. M., "Introduction to Fuzzy Arithmetic: Theory and Applications." Van Nostrand Reinhold, New York, 1991.
11. Bowles, J. B., "Applying Fuzzy Logic to the Design Process." Annual Reliability and Maintainability Symposium. Las Vegas, Nevada, January, 1996.
12. Bowles, J. B. and Palaez, C. E., "Application of Fuzzy Logic to Reliability Engineering." Proc. IEEE, Vol. 83, No. 3, p. 435-449, March 1995.

APPENDIX

Reliability Data for Capacitors

| Part Desc. | Quality Level | App Data Env Source | Part Characteristics | Fail/Hours or Miles (E6) |
|--|---------------|---------------------|---|--------------------------|
| Capacitor, Fixed, Electrolytic, Tantalum | | | | |
| Commercial | GB | 13567-021 | -P#:T110D476M035AS,Mfr:Various,Capacitance Va:47.000u, Voltage Rating:35.000v,Pop:44384 | 0/57.6992 |
| | | | -P#:T110C106M050AS,Mfr:Various,Capacitance Va:10.000u, Voltage Rating:50.000v,Pop:12392 | 0/16.1096 |
| | | | -UP#:0180-3812,Mfr:Various,Capacitance Va:15.000u,Voltage Rating:20.000v, Pop:100344 | 0/130.4472 |
| | | | -UP#:0180-3813,Mfr:Various,Capacitance Va:10.000u,Voltage Rating:10.000v, Pop:31072 | 0/40.3936 |
| | | | -UP#:0180-3818,Mfr:Various,Capacitance Va:4.700u,Voltage Rating:25.000v, Pop:344 | 0/0.4472 |
| | | | -UP#:0180-3822,Mfr:Various,Capacitance Va:39.000u,Voltage Rating:15.000v, Pop:3436 | 0/4.4668 |
| | | | -UP#:0180-3827,Mfr:Various,Capacitance Va:3.300u,Voltage Rating:25.000v, Pop:79620 | 0/103.5060 |
| | | | -UP#:0180-3831,Mfr:Various,Capacitance Va:10.000u,Voltage Rating:35.000v, Pop:323552 | 0/420.6176 |
| | | | -UP#:0180-3833,Mfr:Various,Capacitance Va:22.000u,Voltage Rating:10.000v, Pop:6056 | 0/7.8728 |
| | | | -UP#:0180-3834,Mfr:Various,Capacitance Va:33.000u,Voltage Rating:10.000v, Pop:81648 | 0/106.1424 |
| | | | -UP#:0180-3841,Mfr:Various,Capacitance Va:4.700u,Voltage Rating:16.000v, Pop:260 | 0/0.3380 |
| | | | -UP#:0180-3844,Mfr:Various,Capacitance Va:3.300u,Voltage Rating:35.000v, Pop:20960 | 0/27.2480 |
| | | | -UP#:0180-3845,Mfr:Various,Capacitance Va:4.700u,Voltage Rating:35.000v, Pop:38272 | 0/49.7536 |
| | | | -UP#:0180-3846,Mfr:Various,Capacitance Va:15.000u,Voltage Rating:25.000v, Pop:19264 | 0/25.0432 |
| | | | -UP#:0180-3847,Mfr:Various,Capacitance Va:22.000u,Voltage Rating:25.000v, Pop:31600 | 0/41.0800 |
| | | | -UP#:0180-3848,Mfr:Various,Capacitance Va:33.000u,Voltage Rating:16.000v, Pop:44528 | 0/57.8864 |
| | | | -UP#:0180-3849,Mfr:Various,Capacitance Va:47.000u,Voltage Rating:10.000v, Pop:26440 | 0/34.3720 |
| | | | -UP#:0180-3850,Mfr:Various,Capacitance Va:68.000u,Voltage Rating:10.000v, Pop:36500 | 4/47.4500 |
| | | | -P#:195D224X0035S3,Mfr:3M Co.,Capacitance Va:220u,Voltage Rating:35.000v, Pop:2472 | 0/3.2136 |
| | | | -P#:T356H226M025AS,Mfr:Kemet Electronics Corp.,Capacitance Va:22.000u, Voltage Rating:25.000v,Pop:31796 | 0/41.3348 |
| | | | -UP#:0180-3887,Mfr:Various,Capacitance Va:15.000u,Voltage Rating:20.000v, Pop:44 | 0/0.0572 |
| | | | -UP#:0180-3888,Mfr:Various,Capacitance Va:15.000u,Voltage Rating:25.000v, Pop:5184 | 0/6.7392 |
| | | | -UP#:0180-3889,Mfr:Various,Capacitance Va:6.800u,Voltage Rating:35.000v, Pop:2592 | 0/3.3696 |
| | | | -UP#:0180-3922,Mfr:Various,Capacitance Va:15.000u,Voltage Rating:20.000v, Pop:86968 | 0/113.0584 |
| | | | -P#:T396J107K010AS,Mfr:Kemet Electronics Corp.,Capacitance Va:100.000u, Voltage Rating:10.000v,Pop:9824 | 0/12.7712 |
| | | | -P#:195D105X0035S3,Mfr:3M Co.,Capacitance Va:1.000u,Voltage Rating:35.000v, Pop:824 | 0/1.0712 |
| | | | -UP#:0180-3985,Mfr:Various,Capacitance Va:47.000u,Voltage Rating:6.300v, Pop:1188 | 0/1.5444 |
| | | | -P#:T398E106K025AS,Mfr:Kemet Electronics Corp.,Capacitance Va:10.000u, Voltage Rating:25.000v,Pop:68000 | 0/88.4000 |
| | | | -UP#:0180-4015,Mfr:Various,Capacitance Va:22.000u,Voltage Rating:16.000v, Pop:13868 | 0/18.0284 |
| | | | -P#:T398G475K050AS,Mfr:Kemet Electronics Corp.,Capacitance Va:4.700u, Voltage Rating:50.000v,Pop:20048 | 0/26.0624 |
| | | | -UP#:0180-4031,Mfr:Various,Capacitance Va:10.000u,Voltage Rating:35.000v, Pop:232 | 0/0.3016 |
| | | | -P#:T398B105K050AS,Mfr:Kemet Electronics Corp.,Capacitance Va:1.000u, Voltage Rating:50.000v,Pop:260 | 0/0.3380 |
| | | | -UP#:0180-0022,Mfr:Various,Capacitance Va:3.900u,Voltage Rating:35.000v, Pop:9216 | 0/11.9808 |
| | | | -UP#:0180-0097,Mfr:Various,Capacitance Va:47.000u,Voltage Rating:35.000v, Pop:211596 | 0/275.0748 |
| | | | -UP#:0180-0098,Mfr:Various,Capacitance Va:100.000u,Voltage Rating:20.000v, Pop:107136 | 0/139.2768 |
| | | | -UP#:0180-0100,Mfr:Various,Capacitance Va:4.700u,Voltage Rating:35.000v, Pop:863596 | 4/1122.6748 |
| | | | -UP#:0180-0106,Mfr:Various,Capacitance Va:60.000u,Voltage Rating:6.000v, Pop:224328 | 0/291.6264 |
| | | | -UP#:0180-0113,Mfr:Various,Style:Wet Slug,Capacitance Va:100.000u, Voltage Rating:30.000v,Pop:5044 | 0/6.5572 |
| | | | -UP#:0180-0116,Mfr:Various,Capacitance Va:6.800u,Voltage Rating:35.000v, Pop:2675700 | 16/3478.4100 |
| | | | -UP#:0180-0117,Mfr:Various,Capacitance Va:2.700u,Voltage Rating:35.000v, Pop:29852 | 0/38.8076 |
| | | | -UP#:0180-0137,Mfr:Various,Capacitance Va:100.000u,Voltage Rating:10.000v, Pop:30580 | 0/39.7540 |
| | | | -UP#:0180-0155,Mfr:Various,Capacitance Va:2.200u,Voltage Rating:20.000v, Pop:205212 | 8/266.7756 |
| | | | -UP#:0180-0159,Mfr:Various,Capacitance Va:220.000u,Voltage Rating:10.000v, Pop:84164 | 0/109.4132 |
| | | | -UP#:0180-0160,Mfr:Various,Capacitance Va:22.000u,Voltage Rating:35.000v, Pop:80524 | 0/104.6812 |
| | | | -UP#:0180-0161,Mfr:Various,Capacitance Va:3.300u,Voltage Rating:35.000v, Pop:79016 | 0/102.7208 |
| | | | -UP#:0180-0194,Mfr:Various,Capacitance Va:150.000u,Voltage Rating:15.000v, Pop:8416 | 0/10.9408 |
| | | | -UP#:0180-0195,Mfr:Various,Capacitance Va:330u,Voltage Rating:35.000v, Pop:25200 | 0/32.7600 |
| | | | -UP#:0180-0196,Mfr:Various,Capacitance Va:56.000u,Voltage Rating:15.000v, Pop:17972 | 0/23.3636 |
| | | | -UP#:0180-0197,Mfr:Various,Capacitance Va:2.200u,Voltage Rating:20.000v, Pop:3239896 | 4/4211.8648 |
| | | | -UP#:0180-0210,Mfr:Various,Capacitance Va:3.300u,Voltage Rating:15.000v, Pop:444988 | 4/578.4844 |
| | | | -UP#:0180-0216,Mfr:Various,Capacitance Va:12.000u,Voltage Rating:35.000v, Pop:792 | 0/1.0296 |

| Part Desc. | Quality Level | App Data Env Source | Part Characteristics | Fail/Hours or Miles (E6) |
|--|---------------|---------------------|--|--------------------------|
| Capacitor, Fixed, Electrolytic, Tantalum | | | | |
| Commercial | GB | 13567-021 | UP# 0180-0630, Mfr: Various, Capacitance Va: 4.700u, Voltage Rating: 50.000v, Pop: 100920 | 0/131.1960 |
| | | | -UP# 0180-0642, Mfr: Various, Capacitance Va: 15.000u, Voltage Rating: 20.000v, Pop: 12644 | 0/16.4372 |
| | | | -UP# 0180-0648, Mfr: Various, Capacitance Va: .100u, Voltage Rating: 35.000v, Pop: 11424 | 0/14.8512 |
| | | | -UP# 0180-0678, Mfr: Various, Capacitance Va: 220.000u, Voltage Rating: 3.000v, Pop: 12712 | 0/16.5256 |
| | | | -P# MZG-015-825R-10, Mfr: AVX Corp., Capacitance Va: 8.200u, Voltage Rating: 15.000v, Pop: 12712 | 0/16.5256 |
| | | | -P# MZA-006-685-R20, Mfr: AVX Corp., Capacitance Va: 6.800u, Voltage Rating: 6.000v, Pop: 108 | 0/0.1404 |
| | | | -UP# 0180-0690, Mfr: Various, Capacitance Va: 2.200u, Voltage Rating: 20.000v, Pop: 8 | 0/0.0104 |
| | | | -UP# 0180-1701, Mfr: Various, Capacitance Va: 6.800u, Voltage Rating: 6.000v, Pop: 189440 | 0/246.2720 |
| | | | -UP# 0180-1702, Mfr: Various, Capacitance Va: 180.000u, Voltage Rating: 6.000v, Pop: 8472 | 0/11.0136 |
| | | | -UP# 0180-1704, Mfr: Various, Capacitance Va: 47.000u, Voltage Rating: 6.000v, Pop: 125896 | 0/163.6648 |
| | | | -UP# 0180-1706, Mfr: Various, Style: Wet Slug, Capacitance Va: 100.000u, Voltage Rating: 25.000v, Pop: 22276 | 16/28.9588 |
| | | | -UP# 0180-1713, Mfr: Various, Capacitance Va: .560u, Voltage Rating: 35.000v, Pop: 3112 | 0/4.0456 |
| | | | -UP# 0180-1714, Mfr: Various, Capacitance Va: 330.000u, Voltage Rating: 6.000v, Pop: 60444 | 0/78.5772 |
| | | | -UP# 0180-1715, Mfr: Various, Capacitance Va: 150.000u, Voltage Rating: 6.000v, Pop: 30648 | 0/39.8424 |
| | | | -UP# 0180-1718, Mfr: Various, Capacitance Va: 56.000u, Voltage Rating: 20.000v, Pop: 3112 | 0/4.0456 |
| | | | -UP# 0180-1731, Mfr: Various, Capacitance Va: 4.700u, Voltage Rating: 50.000v, Pop: 391452 | 8/508.8876 |
| | | | -UP# 0180-1735, Mfr: Various, Capacitance Va: .220u, Voltage Rating: 35.000v, Pop: 223108 | 4/290.0404 |
| | | | -UP# 0180-1743, Mfr: Various, Capacitance Va: .100u, Voltage Rating: 35.000v, Pop: 150736 | 0/195.9568 |
| | | | -UP# 0180-1745, Mfr: Various, Capacitance Va: 1.500u, Voltage Rating: 20.000v, Pop: 140872 | 0/183.1336 |
| | | | -UP# 0180-1746, Mfr: Various, Capacitance Va: 15.000u, Voltage Rating: 20.000v, Pop: 2551404 | 36/3316.8252 |
| | | | -UP# 0180-1747, Mfr: Various, Capacitance Va: 150.000u, Voltage Rating: 15.000v, Pop: 14856 | 0/19.3128 |
| | | | -UP# 0180-1759, Mfr: Various, Capacitance Va: .056u, Voltage Rating: 35.000v, Pop: 100 | 0/0.1300 |
| | | | -UP# 0180-1760, Mfr: Various, Capacitance Va: .120u, Voltage Rating: 35.000v, Pop: 100 | 0/0.1300 |
| | | | -UP# 0180-1761, Mfr: Various, Capacitance Va: .560u, Voltage Rating: 35.000v, Pop: 39624 | 0/51.5112 |
| | | | -UP# 0180-1775, Mfr: Various, Capacitance Va: 1.800u, Voltage Rating: 35.000v, Pop: 1316 | 0/1.7108 |
| | | | -UP# 0180-1779, Mfr: Various, Capacitance Va: 18.000u, Voltage Rating: 35.000v, Pop: 15264 | 0/19.8432 |
| | | | -UP# 0180-1794, Mfr: Various, Capacitance Va: 22.000u, Voltage Rating: 35.000v, Pop: 228388 | 0/296.9044 |
| | | | -UP# 0180-1815, Mfr: Various, Capacitance Va: 7.500u, Voltage Rating: 20.000v, Pop: 4848 | 0/6.3024 |
| | | | -UP# 0180-1834, Mfr: Various, Capacitance Va: 15.000u, Voltage Rating: 50.000v, Pop: 28024 | 0/36.4312 |
| | | | -UP# 0180-1835, Mfr: Various, Capacitance Va: 68.000u, Voltage Rating: 15.000v, Pop: 24172 | 0/31.4236 |
| | | | -UP# 0180-1846, Mfr: Various, Capacitance Va: 2.200u, Voltage Rating: 35.000v, Pop: 119740 | 0/155.6620 |
| | | | -UP# 0180-1861, Mfr: Various, Capacitance Va: 27.000u, Voltage Rating: 10.000v, Pop: 53184 | 0/69.1392 |
| | | | -UP# 0180-1940, Mfr: Various, Style: Wet Slug, Capacitance Va: 33.000u, Voltage Rating: 10.000v, Pop: 3752 | 0/4.8776 |
| | | | -UP# 0180-1954, Mfr: Various, Capacitance Va: 4.700u, Voltage Rating: 6.000v, Pop: 7072 | 0/9.1936 |
| | | | -UP# 0180-1974, Mfr: Various, Capacitance Va: 10.000u, Voltage Rating: 35.000v, Pop: 44280 | 0/57.5640 |
| | | | -UP# 0180-1980, Mfr: Various, Capacitance Va: 1.000u, Voltage Rating: 35.000v, Pop: 29388 | 0/38.2044 |
| | | | -UP# 0180-2050, Mfr: Various, Capacitance Va: .082u, Voltage Rating: 35.000v, Pop: 5860 | 0/7.6180 |
| | | | -UP# 0180-2060, Mfr: Various, Capacitance Va: 40.000u, Voltage Rating: 10.000v, Pop: 19252 | 0/25.0276 |
| | | | -UP# 0180-2062, Mfr: Various, Capacitance Va: 120.000u, Voltage Rating: 10.000v, Pop: 5004 | 0/6.5052 |
| | | | -UP# 0180-2071, Mfr: Various, Capacitance Va: .022u, Voltage Rating: 35.000v, Pop: 9280 | 0/12.0640 |
| | | | -UP# 0180-2076, Mfr: Various, Capacitance Va: .068u, Voltage Rating: 35.000v, Pop: 36 | 0/0.0468 |
| | | | -UP# 0180-2079, Mfr: Various, Capacitance Va: .390u, Voltage Rating: 35.000v, Pop: 592 | 0/0.7696 |
| | | | -UP# 0180-2104, Mfr: Various, Style: Wet Slug, Capacitance Va: 40.000u, Voltage Rating: 30.000v, Pop: 3488 | 0/4.5344 |
| | | | -UP# 0180-2111, Mfr: Various, Capacitance Va: 33.000u, Voltage Rating: 35.000v, Pop: 7200 | 0/9.3600 |

3-116 Part Details

EPRD-97

| Part Desc. | Quality Level | App Data Env Source | Part Characteristics | Fail/Hours or Miles (E6) |
|--|---------------|---------------------|---|--------------------------|
| Capacitor, Fixed, Electrolytic, Tantalum | | | | |
| Commercial | GB | 13567-021 | UP#0180-0218, Mfr: Various, Capacitance Va: .150u, Voltage Rating: 35.000v, Pop: 39004 | 0/50.7052 |
| | | | -UP#0180-0228, Mfr: Various, Capacitance Va: 22.000u, Voltage Rating: 15.000v, Pop: 2047428 | 8/2661.6564 |
| | | | -UP#0180-0229, Mfr: Various, Capacitance Va: 33.000u, Voltage Rating: 10.000v, Pop: 1930224 | 4/2509.2912 |
| | | | -UP#0180-0230, Mfr: Various, Capacitance Va: 1.000u, Voltage Rating: 50.000v, Pop: 574664 | 12/747.0632 |
| | | | -UP#0180-0232, Mfr: Various, Style: Wet Slug, Capacitance Va: 10.000u, Voltage Rating: 100.000v, Pop: 388 | 0/0.5044 |
| | | | -UP#0180-0234, Mfr: Various, Style: Wet Slug, Capacitance Va: 33.000u, Voltage Rating: 75.000v, Pop: 16256 | 0/21.1328 |
| | | | -UP#0180-0235, Mfr: Various, Style: Wet Slug, Capacitance Va: 56.000u, Voltage Rating: 75.000v, Pop: 2748 | 0/3.5724 |
| | | | -UP#0180-0291, Mfr: Various, Capacitance Va: 1.000u, Voltage Rating: 35.000v, Pop: 4220396 | 16/5486.5148 |
| | | | -UP#0180-0309, Mfr: Various, Capacitance Va: 4.700u, Voltage Rating: 10.000v, Pop: 661456 | 0/859.8928 |
| | | | -UP#0180-0347, Mfr: Various, Capacitance Va: 1.500u, Voltage Rating: 35.000v, Pop: 12340 | 0/16.0420 |
| | | | -UP#0180-0349, Mfr: Various, Capacitance Va: .820u, Voltage Rating: 35.000v, Pop: 7332 | 0/9.5316 |
| | | | -UP#0180-0354, Mfr: Various, Capacitance Va: 40.000u, Voltage Rating: 10.000v, Pop: 5940 | 0/7.7220 |
| | | | -P#151D145X0035X2, Mfr: 3M Co., Capacitance Va: 3.400u, Voltage Rating: 35.000v, Pop: 2456 | 0/3.1928 |
| | | | -UP#0180-0356, Mfr: Various, Style: Wet Slug, Capacitance Va: 70.000u, Voltage Rating: 15.000v, Pop: 448 | 0/0.5824 |
| | | | -UP#0180-0373, Mfr: Various, Capacitance Va: .680u, Voltage Rating: 35.000v, Pop: 205040 | 0/266.5520 |
| | | | -UP#0180-0374, Mfr: Various, Capacitance Va: 10.000u, Voltage Rating: 20.000v, Pop: 2314796 | 20/3009.2348 |
| | | | -UP#0180-0375, Mfr: Various, Capacitance Va: 68.000u, Voltage Rating: 20.000v, Pop: 39308 | 0/51.1004 |
| | | | -UP#0180-0376, Mfr: Various, Capacitance Va: .470u, Voltage Rating: 35.000v, Pop: 269584 | 0/350.4592 |
| | | | -UP#0180-0387, Mfr: Various, Capacitance Va: 47.000u, Voltage Rating: 20.000v, Pop: 4180 | 0/5.4340 |
| | | | -UP#0180-0388, Mfr: Various, Style: Wet Slug, Capacitance Va: 56.000u, Voltage Rating: 20.000v, Pop: 3896 | 0/5.0648 |
| | | | -UP#0180-0393, Mfr: Various, Capacitance Va: 39.000u, Voltage Rating: 10.000v, Pop: 64564 | 0/83.9332 |
| | | | -UP#0180-0405, Mfr: Various, Capacitance Va: 1.800u, Voltage Rating: 20.000v, Pop: 49976 | 0/64.9688 |
| | | | -UP#0180-0407, Mfr: Various, Style: Wet Slug, Capacitance Va: 180.000u, Voltage Rating: 30.000v, Pop: 12208 | 0/15.8704 |
| | | | -UP#0180-0415, Mfr: Various, Capacitance Va: 10.000u, Voltage Rating: 25.000v, Pop: 3560 | 0/4.6280 |
| | | | -UP#0180-0418, Mfr: Various, Capacitance Va: 1.000u, Voltage Rating: 35.000v, Pop: 89824 | 4/116.7712 |
| | | | -UP#0180-0428, Mfr: Various, Capacitance Va: 68.000u, Voltage Rating: 6.000v, Pop: 15012 | 0/19.5156 |
| | | | -UP#0180-0437, Mfr: Various, Capacitance Va: 39.000u, Voltage Rating: 10.000v, Pop: 63144 | 0/82.0872 |
| | | | -UP#0180-0451, Mfr: Various, Capacitance Va: 3.300u, Voltage Rating: 20.000v, Pop: 41340 | 0/53.7420 |
| | | | -P#T376C105M020AS, Mfr: Kemet Electronics Corp., Capacitance Va: 1.000u, Voltage Rating: 20.000v, Pop: 8 | 0/0.0104 |
| | | | -UP#0180-0474, Mfr: Various, Capacitance Va: 15.000u, Voltage Rating: 20.000v, Pop: 150316 | 0/195.4108 |
| | | | -UP#0180-0479, Mfr: Various, Capacitance Va: 47.000u, Voltage Rating: 25.000v, Pop: 9792 | 0/12.7296 |
| | | | -UP#0180-0481, Mfr: Various, Style: Wet Slug, Capacitance Va: 100.000u, Voltage Rating: 20.000v, Pop: 3016 | 0/3.9208 |
| | | | -UP#0180-0486, Mfr: Various, Capacitance Va: 10.000u, Voltage Rating: 20.000v, Pop: 2372 | 0/3.0836 |
| | | | -UP#0180-0490, Mfr: Various, Capacitance Va: 68.000u, Voltage Rating: 6.000v, Pop: 36832 | 0/47.8816 |
| | | | -UP#0180-0491, Mfr: Various, Capacitance Va: 10.000u, Voltage Rating: 25.000v, Pop: 686180 | 0/892.0340 |
| | | | -UP#0180-0500, Mfr: Various, Capacitance Va: 47.000u, Voltage Rating: 20.000v, Pop: 49760 | 0/64.6880 |
| | | | -UP#0180-0548, Mfr: Various, Capacitance Va: 56.000u, Voltage Rating: 6.000v, Pop: 4 | 0/0.0052 |
| | | | -UP#0180-0552, Mfr: Various, Capacitance Va: 220.000u, Voltage Rating: 10.000v, Pop: 41956 | 0/54.5428 |
| | | | -UP#0180-0553, Mfr: Various, Capacitance Va: 22.000u, Voltage Rating: 25.000v, Pop: 179028 | 4/232.7364 |
| | | | -UP#0180-0558, Mfr: Various, Capacitance Va: 470.000u, Voltage Rating: 10.000v, Pop: 4364 | 0/5.6732 |
| | | | -UP#0180-0562, Mfr: Various, Capacitance Va: 33.000u, Voltage Rating: 10.000v, Pop: 183348 | 0/238.3524 |
| | | | -UP#0180-0575, Mfr: Various, Capacitance Va: 2.200u, Voltage Rating: 15.000v, Pop: 1800 | 0/2.3400 |
| | | | -UP#0180-0594, Mfr: Various, Capacitance Va: 3.300u, Voltage Rating: 15.000v, Pop: 63276 | 0/82.2588 |
| | | | -UP#0180-0597, Mfr: Various, Capacitance Va: 22.000u, Voltage Rating: 50.000v, Pop: 2180 | 0/2.8340 |

3-118 Part Details

EPRD-97

| Part Desc. | Quality Level | App Data Env Source | Part Characteristics | Fail/Hours or Miles (E6) |
|--|---------------|---------------------|--|--------------------------|
| Capacitor, Fixed, Electrolytic, Tantalum | | | | |
| Commercial | GB | 13567-021 | -UP# 0180-2125, Mfr: Various, Capacitance Va: 15.000u, Voltage Rating: 20.000v, Pop: 2248 | 0/2.9224 |
| | | | -UP# 0180-2126, Mfr: Various, Capacitance Va: 1.500u, Voltage Rating: 35.000v, Pop: 8976 | 0/11.6688 |
| | | | -UP# 0180-2127, Mfr: Various, Capacitance Va: .150u, Voltage Rating: 35.000v, Pop: 9244 | 0/12.0172 |
| | | | -UP# 0180-2129, Mfr: Various, Capacitance Va: 10.000u, Voltage Rating: 50.000v, Pop: 19704 | 0/25.6152 |
| | | | -P# 109D106X0060C2-DYP, Mfr: Sprague Electric Co., Style: Wet Slug, Capacitance Va: 10.000u, Voltage Rating: 60.000v, Pop: 47312 | 4/61.5056 |
| | | | -UP# 0180-2140, Mfr: Various, Capacitance Va: 5.600u, Voltage Rating: 50.000v, Pop: 12720 | 0/16.5360 |
| | | | -UP# 0180-2141, Mfr: Various, Capacitance Va: 3.300u, Voltage Rating: 50.000v, Pop: 158952 | 4/206.6376 |
| | | | -UP# 0180-2148, Mfr: Various, Capacitance Va: .470u, Voltage Rating: 50.000v, Pop: 6368 | 0/8.2784 |
| | | | -UP# 0180-2161, Mfr: Various, Capacitance Va: .750u, Voltage Rating: 50.000v, Pop: 16040 | 0/20.8520 |
| | | | -UP# 0180-2178, Mfr: Various, Style: Wet Slug, Capacitance Va: 220.000u, Voltage Rating: 8.000v, Pop: 3220 | 0/4.1860 |
| | | | -UP# 0180-2182, Mfr: Various, Capacitance Va: 18.000u, Voltage Rating: 50.000v, Pop: 14268 | 0/18.5484 |
| | | | -UP# 0180-2186, Mfr: Various, Style: Wet Slug, Capacitance Va: 300.000u, Voltage Rating: 30.000v, Pop: 9200 | 4/11.9600 |
| | | | -UP# 0180-2192, Mfr: Various, Capacitance Va: 6.800u, Voltage Rating: 50.000v, Pop: 2856 | 0/3.7128 |
| | | | -UP# 0180-2195, Mfr: Various, Capacitance Va: 15.000u, Voltage Rating: 35.000v, Pop: 27916 | 0/36.2908 |
| | | | -UP# 0180-2201, Mfr: Various, Capacitance Va: .680u, Voltage Rating: 75.000v, Pop: 8 | 0/0.0104 |
| | | | -UP# 0180-2204, Mfr: Various, Capacitance Va: 10.000u, Voltage Rating: 10.000v, Pop: 1820 | 0/2.3660 |
| | | | -UP# 0180-2205, Mfr: Various, Capacitance Va: .330u, Voltage Rating: 35.000v, Pop: 180516 | 0/234.6708 |
| | | | -UP# TA ELECTROLYTICB, Mfr: Various, Capacitance Va: 60.000u, Voltage Rating: 6.000v, Pop: 184340 | 0/239.6420 |
| | | | -UP# 0180-2207, Mfr: Various, Capacitance Va: 100.000u, Voltage Rating: 10.000v, Pop: 286280 | 0/372.1640 |
| | | | -UP# 0180-2208, Mfr: Various, Capacitance Va: 220.000u, Voltage Rating: 10.000v, Pop: 161604 | 0/210.0852 |
| | | | -UP# 0180-2247, Mfr: Various, Capacitance Va: 10.000u, Voltage Rating: 20.000v, Pop: 6236 | 0/8.1068 |
| | | | -UP# 0180-2249, Mfr: Various, Capacitance Va: 47.000u, Voltage Rating: 20.000v, Pop: 158224 | 0/205.6912 |
| | | | -UP# 0180-2255, Mfr: Various, Capacitance Va: 2.200u, Voltage Rating: 20.000v, Pop: 53192 | 0/69.1496 |
| | | | -UP# 0180-2264, Mfr: Various, Capacitance Va: 3.300u, Voltage Rating: 15.000v, Pop: 58676 | 0/76.2788 |
| | | | -UP# 0180-2268, Mfr: Various, Style: Wet Slug, Capacitance Va: 140.000u, Voltage Rating: 30.000v, Pop: 1832 | 0/2.3816 |
| | | | -UP# 0180-2275, Mfr: Various, Style: Wet Slug, Capacitance Va: 220.000u, Voltage Rating: 8.000v, Pop: 16936 | 0/22.0168 |
| | | | -UP# 0180-2333, Mfr: Various, Style: Wet Slug, Capacitance Va: 330.000u, Voltage Rating: 15.000v, Pop: 7620 | 0/9.9060 |
| | | | -UP# 0180-2338, Mfr: Various, Style: Wet Slug, Capacitance Va: 650.000u, Voltage Rating: 13.000v, Pop: 11000 | 0/14.3000 |
| | | | -UP# 0180-2358, Mfr: Various, Capacitance Va: 150.000u, Voltage Rating: 6.000v, Pop: 4 | 0/0.0052 |
| | | | -UP# 0180-2374, Mfr: Various, Capacitance Va: 100.000u, Voltage Rating: 20.000v, Pop: 19000 | 0/24.7000 |
| | | | -P# T370C225K025AS, Mfr: Kemet Electronics Corp., Capacitance Va: 2.200u, Voltage Rating: 25.000v, Pop: 20588 | 0/26.7644 |
| | | | -UP# 0180-2378, Mfr: Various, Capacitance Va: 6.800u, Voltage Rating: 10.000v, Pop: 41176 | 0/53.5288 |
| | | | -UP# 0180-2419, Mfr: Various, Style: Wet Slug, Capacitance Va: 470.000u, Voltage Rating: 10.000v, Pop: 2952 | 0/3.8376 |
| | | | -UP# 0180-2474, Mfr: Various, Style: Wet Slug, Capacitance Va: 15.000u, Voltage Rating: 20.000v, Pop: 20 | 0/0.0260 |
| | | | -UP# 0180-2486, Mfr: Various, Style: Wet Slug, Capacitance Va: 470.000u, Voltage Rating: 30.000v, Pop: 14224 | 0/18.4912 |
| | | | -UP# 0180-2505, Mfr: Various, Capacitance Va: 1.000u, Voltage Rating: 75.000v, Pop: 35284 | 0/45.8692 |
| | | | -P# MMF-002-106A-20, Mfr: AVX Corp., Capacitance Va: 10.000u, Voltage Rating: 2.000v, Pop: 32 | 0/0.0416 |
| | | | -UP# 0180-2514, Mfr: Various, Capacitance Va: 2.200u, Voltage Rating: 20.000v, Pop: 2208 | 0/2.8704 |
| | | | -UP# 0180-2515, Mfr: Various, Capacitance Va: 47.000u, Voltage Rating: 6.000v, Pop: 88028 | 0/114.4364 |
| | | | -UP# 0180-2545, Mfr: Various, Capacitance Va: 100.000u, Voltage Rating: 4.000v, Pop: 50564 | 0/65.7332 |
| | | | -UP# 0180-2549, Mfr: Various, Capacitance Va: 100.000u, Voltage Rating: 20.000v, Pop: 528 | 0/0.6864 |
| | | | -UP# 0180-2597, Mfr: Various, Style: Wet Slug, Capacitance Va: 270.000u, Voltage Rating: 25.000v, Pop: 8952 | 0/11.6376 |
| | | | -P# MD7-008-476-20/9038, Mfr: AVX Corp., Capacitance Va: 47.000u, Voltage Rating: 8.000v, Pop: 7928 | 0/10.3064 |
| | | | -P# MAZ-035-684-R20, Mfr: AVX Corp., Capacitance Va: .680u, Voltage Rating: 35.000v, Pop: 12 | 0/0.0156 |
| | | | -UP# 0180-2610, Mfr: Various, Capacitance Va: 10.000u, Voltage Rating: 75.000v, Pop: 18780 | 0/24.4140 |

| Part Desc. | Quality Level | App Data Env Source | Part Characteristics | Fail/Hours or Miles (E6) |
|--|---------------|---------------------|--|--------------------------|
| Capacitor, Fixed, Electrolytic, Tantalum | | | | |
| Commercial | GB | 13567-021 | UP#:0180-2613,Mfr:Various,Capacitance Va:390.000u,Voltage Rating:6.000v,Pop:4728 | 0/6.1464 |
| | | | -UP#:0180-2614,Mfr:Various,Capacitance Va:100.000u,Voltage Rating:30.000v,Pop:50940 | 0/66.2220 |
| | | | -P#:KD7-010-226-20/9038,Mfr:AVX Corp.,Capacitance Va:22.000u,Voltage Rating:10.000v,Pop:6148 | 0/7.9924 |
| | | | -UP#:0180-2616,Mfr:Various,Capacitance Va:60.000u,Voltage Rating:6.000v,Pop:10532 | 0/13.6916 |
| | | | -UP#:0180-2617,Mfr:Various,Capacitance Va:6.800u,Voltage Rating:35.000v,Pop:344136 | 0/447.3768 |
| | | | -UP#:0180-2618,Mfr:Various,Capacitance Va:33.000u,Voltage Rating:10.000v,Pop:172656 | 0/224.4528 |
| | | | -UP#:0180-2619,Mfr:Various,Capacitance Va:22.000u,Voltage Rating:15.000v,Pop:138856 | 0/180.5128 |
| | | | -UP#:0180-2620,Mfr:Various,Capacitance Va:2.200u,Voltage Rating:50.000v,Pop:326404 | 12/424.3252 |
| | | | -UP#:0180-2623,Mfr:Various,Capacitance Va:12.000u,Voltage Rating:6.000v,Pop:55176 | 0/71.7288 |
| | | | -UP#:0180-2661,Mfr:Various,Capacitance Va:1.000u,Voltage Rating:50.000v,Pop:189368 | 0/246.1784 |
| | | | -UP#:0180-2662,Mfr:Various,Capacitance Va:10.000u,Voltage Rating:10.000v,Pop:113368 | 0/147.3784 |
| | | | -UP#:0180-2664,Mfr:Various,Capacitance Va:3.300u,Voltage Rating:15.000v,Pop:1172 | 0/1.5236 |
| | | | -UP#:0180-2667,Mfr:Various,Capacitance Va:150.000u,Voltage Rating:20.000v,Pop:38076 | 0/49.4988 |
| | | | -P#:KD6-035-475-20/9038,Mfr:AVX Corp.,Capacitance Va:4.700u,Voltage Rating:35.000v,Pop:66856 | 0/86.9128 |
| | | | -P#:KD7-008-336-20/9038,Mfr:AVX Corp.,Capacitance Va:33.000u,Voltage Rating:8.000v,Pop:1892 | 0/2.4596 |
| | | | -UP#:0180-2688,Mfr:Various,Capacitance Va:8.200u,Voltage Rating:20.000v,Pop:12 | 0/0.0156 |
| | | | -UP#:0180-2690,Mfr:Various,Capacitance Va:3.300u,Voltage Rating:15.000v,Pop:176392 | 4/229.3096 |
| | | | -UP#:0180-2692,Mfr:Various,Capacitance Va:68.000u,Voltage Rating:15.000v,Pop:4812 | 0/6.2556 |
| | | | -UP#:0180-2697,Mfr:Various,Capacitance Va:10.000u,Voltage Rating:25.000v,Pop:341376 | 4/443.7888 |
| | | | -UP#:0180-2698,Mfr:Various,Capacitance Va:4.700u,Voltage Rating:35.000v,Pop:28696 | 0/37.3048 |
| | | | -P#:MS-020-225-10/9038,Mfr:AVX Corp.,Capacitance Va:2.200u,Voltage Rating:20.000v,Pop:30516 | 0/39.6708 |
| | | | -UP#:0180-2743,Mfr:Various,Capacitance Va:110u,Voltage Rating:35.000v,Pop:1128 | 0/1.4664 |
| | | | -UP#:0180-2764,Mfr:Various,Capacitance Va:1.000u,Voltage Rating:35.000v,Pop:1700 | 0/2.2100 |
| | | | -P#:KD7-020-156-20/9038,Mfr:AVX Corp.,Capacitance Va:15.000u,Voltage Rating:20.000v,Pop:33436 | 0/43.4668 |
| | | | -UP#:0180-2781,Mfr:Various,Capacitance Va:39.000u,Voltage Rating:10.000v,Pop:12828 | 0/16.6764 |
| | | | -UP#:0180-2794,Mfr:Various,Capacitance Va:3.300u,Voltage Rating:35.000v,Pop:6960 | 0/9.0480 |
| | | | -UP#:0180-2795,Mfr:Various,Capacitance Va:39.000u,Voltage Rating:15.000v,Pop:7480 | 0/9.7240 |
| | | | -UP#:0180-2811,Mfr:Various,Capacitance Va:10.000u,Voltage Rating:35.000v,Pop:129120 | 0/167.8560 |
| | | | -P#:394D684X0025B3,Mfr:3M Co.,Capacitance Va:680u,Voltage Rating:25.000v,Pop:8960 | 0/11.6480 |
| | | | -UP#:0180-2814,Mfr:Various,Capacitance Va:22.000u,Voltage Rating:10.000v,Pop:87336 | 0/113.5368 |
| | | | -UP#:0180-2815,Mfr:Various,Capacitance Va:100.000u,Voltage Rating:10.000v,Pop:170696 | 0/221.9048 |
| | | | -UP#:0180-2816,Mfr:Various,Capacitance Va:68.000u,Voltage Rating:10.000v,Pop:61604 | 0/80.0852 |
| | | | -UP#:0180-2817,Mfr:Various,Capacitance Va:47.000u,Voltage Rating:10.000v,Pop:43056 | 0/55.9728 |
| | | | -UP#:0180-2818,Mfr:Various,Capacitance Va:2.200u,Voltage Rating:35.000v,Pop:227612 | 8/295.8956 |
| | | | -UP#:0180-2819,Mfr:Various,Capacitance Va:470u,Voltage Rating:35.000v,Pop:12580 | 0/16.3540 |
| | | | -UP#:0180-2820,Mfr:Various,Capacitance Va:220u,Voltage Rating:35.000v,Pop:22152 | 0/28.7976 |
| | | | -UP#:0180-2821,Mfr:Various,Capacitance Va:22.000u,Voltage Rating:35.000v,Pop:191384 | 0/248.7992 |
| | | | -P#:550D476X0035S2,Mfr:3M Co.,Capacitance Va:47.000u,Voltage Rating:35.000v,Pop:1760 | 0/2.2880 |
| | | | -P#:109D826X9075F2,Mfr:Sprague Electric Co.,Style:Wet Slug,Capacitance Va:82.000u,Voltage Rating:75.000v,Pop:4360 | 0/5.6680 |
| | | | -P#:109D106X0100C2,Mfr:Sprague Electric Co.,Style:Wet Slug,Capacitance Va:10.000u,Voltage Rating:100.000v,Pop:4088 | 0/5.3144 |
| | | | -P#:109D686X9100T2,Mfr:Sprague Electric Co.,Style:Wet Slug,Capacitance Va:68.000u,Voltage Rating:100.000v,Pop:4700 | 0/6.1100 |
| | | | -UP#:0180-2900,Mfr:Various,Capacitance Va:15.000u,Voltage Rating:75.000v,Pop:3536 | 0/4.5968 |
| | | | -UP#:0180-2904,Mfr:Various,Capacitance Va:100u,Voltage Rating:75.000v,Pop:6360 | 0/8.2680 |
| | | | -UP#:0180-2925,Mfr:Various,Capacitance Va:82.000u,Voltage Rating:10.000v,Pop:3884 | 0/5.0492 |

3-120 Part Details

EPRD-97

| Part Desc. | Quality Level | App Data Env Source | Part Characteristics | Fail/Hours or Miles (E6) |
|--|---------------|---------------------|--|--------------------------|
| Capacitor, Fixed, Electrolytic, Tantalum | | | | |
| Commercial | GB | 13567-021 | -UP#:0180-2927,Mfr:Various,Style:Wet Slug,Capacitance Va:650.000u, Voltage Rating:20.000v,Pop:1316 | 0/1.7108 |
| | | | -UP#:0180-2929,Mfr:Various,Capacitance Va:68.000u,Voltage Rating:10.000v, Pop:128252 | 0/166.7276 |
| | | | -UP#:0180-2944,Mfr:Various,Capacitance Va:15.000u,Voltage Rating:20.000v, Pop:12952 | 0/16.8376 |
| | | | -UP#:0180-3018,Mfr:Various,Capacitance Va:4.700u,Voltage Rating:10.000v, Pop:1664 | 0/2.1632 |
| | | | -UP#:0180-3020,Mfr:Various,Style:Wet Slug,Capacitance Va:120.000u, Voltage Rating:50.000v,Pop:21640 | 12/28.1320 |
| | | | -UP#:0180-3051,Mfr:Various,Capacitance Va:150.000u,Voltage Rating:6.000v, Pop:2712 | 0/3.5256 |
| | | | -UP#:0180-3073,Mfr:Various,Capacitance Va:2.200u,Voltage Rating:30.000v, Pop:12688 | 0/16.4944 |
| | | | -UP#:0180-3074,Mfr:Various,Capacitance Va:15.000u,Voltage Rating:30.000v, Pop:24760 | 0/32.1880 |
| | | | -UP#:0180-3084,Mfr:Various,Capacitance Va:22.000u,Voltage Rating:20.000v, Pop:4604 | 0/5.9852 |
| | | | -UP#:0180-3087,Mfr:Various,Capacitance Va:1.500u,Voltage Rating:60.000v, Pop:7112 | 0/9.2456 |
| | | | -UP#:0180-3135,Mfr:Various,Capacitance Va:10.000u,Voltage Rating:10.000v, Pop:16 | 0/0.0208 |
| | | | -UP#:0180-3143,Mfr:Various,Capacitance Va:3.300u,Voltage Rating:75.000v, Pop:20056 | 0/26.0728 |
| | | | -UP#:0180-3210,Mfr:Various,Capacitance Va:27.000u,Voltage Rating:10.000v, Pop:2040 | 0/2.6520 |
| | | | -UP#:0180-3319,Mfr:Various,Capacitance Va:22.000u,Voltage Rating:50.000v, Pop:688 | 0/0.8944 |
| | | | -P#:151D236X903522,Mfr:3M Co.,Capacitance Va:23.000u,Voltage Rating:35.000v, Pop:424 | 0/0.5512 |
| | | | -UP#:0180-3422,Mfr:Various,Capacitance Va:10.000u,Voltage Rating:10.000v, Pop:5176 | 0/6.7288 |
| | | | -UP#:0180-3440,Mfr:Various,Capacitance Va:47.000u,Voltage Rating:10.000v, Pop:12644 | 0/16.4372 |
| | | | -UP#:0180-3487,Mfr:Various,Capacitance Va:150.000u,Voltage Rating:16.000v, Pop:11544 | 0/15.0072 |
| | | | -UP#:0180-3573,Mfr:Various,Capacitance Va:10.000u,Voltage Rating:35.000v, Pop:1984 | 0/2.5792 |
| | | | -UP#:0180-3574,Mfr:Various,Capacitance Va:3.300u,Voltage Rating:30.000v, Pop:2652 | 0/3.4476 |
| | | | -UP#:0180-3629,Mfr:Various,Capacitance Va:15.000u,Voltage Rating:35.000v, Pop:18960 | 0/24.6480 |
| | | | -P#:T368C476M020AS,Mfr:Kemet Electronics Corp.,Capacitance Va:47.000u, Voltage Rating:20.000v,Pop:8932 | 0/11.6116 |
| | | | -P#:196D1164,Mfr:3M Co.,Capacitance Va:2.200u,Voltage Rating:75.000v, Pop:16684 | 0/21.6892 |
| | | | -P#:T350G156M025AS,Mfr:Kemet Electronics Corp.,Capacitance Va:15.000u, Voltage Rating:25.000v,Pop:9968 | 0/12.9584 |
| | | | -UP#:0180-3741,Mfr:Various,Capacitance Va:2.200u,Voltage Rating:25.000v, Pop:16332 | 0/21.2316 |
| | | | -UP#:0180-3743,Mfr:Various,Capacitance Va:.680u,Voltage Rating:50.000v, Pop:18952 | 0/24.6376 |
| | | | -UP#:0180-3744,Mfr:Various,Capacitance Va:4.700u,Voltage Rating:10.000v, Pop:18952 | 0/24.6376 |
| | | | -UP#:0180-3749,Mfr:Various,Capacitance Va:.220u,Voltage Rating:35.000v, Pop:360 | 0/0.4680 |
| | | | -UP#:0180-3750,Mfr:Various,Capacitance Va:4.700u,Voltage Rating:35.000v, Pop:120 | 0/0.1560 |
| | | | -UP#:0180-3751,Mfr:Various,Capacitance Va:1.000u,Voltage Rating:35.000v, Pop:39780 | 0/51.7140 |
| | | | -UP#:0180-3752,Mfr:Various,Capacitance Va:3.300u,Voltage Rating:35.000v, Pop:2868 | 0/3.7284 |
| | | | -UP#:0180-3753,Mfr:Various,Capacitance Va:6.800u,Voltage Rating:35.000v, Pop:33464 | 0/43.5032 |
| | | | -UP#:0180-3754,Mfr:Various,Capacitance Va:10.000u,Voltage Rating:25.000v, Pop:69140 | 0/89.8820 |
| | | | -UP#:0180-3755,Mfr:Various,Capacitance Va:33.000u,Voltage Rating:10.000v, Pop:1536 | 0/1.9968 |
| | | | -UP#:0180-3767,Mfr:Various,Capacitance Va:3.300u,Voltage Rating:25.000v, Pop:14952 | 0/19.4376 |
| | | | -UP#:0180-3768,Mfr:Various,Capacitance Va:3.300u,Voltage Rating:35.000v, Pop:70648 | 0/91.8424 |
| | | | -UP#:0180-3769,Mfr:Various,Capacitance Va:6.800u,Voltage Rating:35.000v, Pop:37704 | 0/49.0152 |
| | | | -UP#:0180-3770,Mfr:Various,Capacitance Va:2.200u,Voltage Rating:35.000v, Pop:108620 | 0/141.2060 |
| | | | -UP#:0180-3771,Mfr:Various,Capacitance Va:1.000u,Voltage Rating:35.000v, Pop:317892 | 0/413.2596 |
| | | | -UP#:0180-3772,Mfr:Various,Capacitance Va:10.000u,Voltage Rating:25.000v, Pop:1728 | 0/2.2464 |
| | | | -UP#:0180-3773,Mfr:Various,Capacitance Va:.470u,Voltage Rating:35.000v, Pop:2860 | 0/3.7180 |
| | | | -UP#:0180-3775,Mfr:Various,Capacitance Va:68.000u,Voltage Rating:10.000v, Pop:167132 | 0/217.2716 |
| | | | -UP#:0180-3784,Mfr:Various,Capacitance Va:22.000u,Voltage Rating:25.000v, Pop:371876 | 0/483.4388 |
| | | | -UP#:0180-3803,Mfr:3M Co.,Capacitance Va:6.800u,Voltage Rating:75.000v, Pop:12000 | 0/15.6000 |

3-130 Part Details

EPRD-97

| Part Desc. | Quality Level | App Data Env Source | Part Characteristics | Fail/Hours or Miles (E6) |
|--|---------------|---------------------|--|--------------------------|
| Capacitor, Fixed, Electrolytic, Tantalum | | | | |
| Military | GP | 14851-000 | -P#:CSR13BC105K, Mil#:M39003-01-2357, NSN:5910-00-495-0042, Mfr:Various, Pkg.:Metal, Herm.:Hermetic, Polarity:Polarized, Tol.:+/-20, Capacitance Va:1.000u, Voltage Rating:50.000D, Pop:1161 | 2/28.8183 |
| | | | -P#:CSR13E475K, Mil#:M39003-01-2368, NSN:5910-00-007-2004, Mfr:Various, Pkg.:Metal, Herm.:Hermetic, Polarity:Polarized, Tol.:+/-10, Capacitance Va:4700000.000p, Voltage Rating:50.000D, Pop:258 | 0/6.4041 |
| | | | -P#:CSR13G475KM, Mil#:M39003-01-2368, NSN:5910-01-086-1108, Mfr:Various, Pkg.:Metal, Herm.:Hermetic, Polarity:Polarized, Tol.:+/-20, Capacitance Va:4700000.000p, Voltage Rating:50.000D, Pop:516 | 0/12.8081 |
| | | | -P#:CSR13G156KM, Mil#:M39003-01-2378, NSN:5910-00-137-7584, Mfr:Various, Pkg.:Metal, Herm.:Hermetic, Polarity:Polarized, Tol.:+/-20, Capacitance Va:15.000u, Voltage Rating:50.000D, Pop:129 | 0/3.2020 |
| | | | -P#:CSR13C685KM, Mil#:M39003-01-3024, NSN:5910-00-137-7584, Mfr:Various, Pkg.:Metal, Herm.:Hermetic, Polarity:Polarized, Tol.:+/-10, Capacitance Va:6800000.000p, Voltage Rating:35.000D, Pop:129 | 0/3.2020 |
| | | | -P#:CSR13C335KM, Mil#:M39003-03-0036, NSN:5910-00-010-8159, Mfr:Various, Style:CSR, Pkg.:Metal, Herm.:Hermetic, Polarity:Polarized, Tol.:+/-20, Capacitance Va:33.000u, Voltage Rating:15.000D, Pop:129 | 0/3.2020 |
| | | | -P#:CSR13C685KM, Mil#:M39003-01-3024, NSN:5910-00-144-4381, Mfr:Various, Style:CSR, Pkg.:Metal, Herm.:Hermetic, Polarity:Polarized, Tol.:+/-10, Capacitance Va:6800000.000p, Voltage Rating:35.000D, Pop:258 | 0/6.4041 |
| | | | -P#:CSR13C475KM, Mil#:M39003-01-2255, NSN:5910-00-416-9775, Mfr:Various, Pkg.:Metal, Polarity:Polarized, Tol.:+/-20, Capacitance Va:4700000.000p, Voltage Rating:10.000D, Pop:129 | 0/3.2020 |
| | | | -P#:CSR13C336KM, Mil#:M39003-01-2257, NSN:5910-00-189-4248, Mfr:Various, Style:CSR, Pkg.:Metal, Herm.:Hermetic, Polarity:Polarized, Tol.:+/-10, Capacitance Va:33.000u, Voltage Rating:10.000D, Pop:129 | 0/3.2020 |
| | | | -P#:CSR13C335KM, Mil#:M39003-01-2259, NSN:5910-00-105-1976, Mfr:Various, Pkg.:Metal, Herm.:Hermetic, Polarity:Polarized, Tol.:+/-10, Capacitance Va:33.000u, Voltage Rating:10.000D, Pop:129 | 0/3.2020 |
| | | | -P#:CSR13C107KM, Mil#:M39003-01-2262, NSN:5910-00-550-1901, Mfr:Various, Pkg.:Metal, Herm.:Hermetic, Polarity:Polarized, Tol.:+/-20, Capacitance Va:100.000u, Voltage Rating:10.000D, Pop:129 | 0/3.2020 |
| | | | -P#:CSR13D226KM, Mil#:M39003-01-2272, NSN:5910-00-165-0629, Mfr:Various, Pkg.:Metal, Herm.:Hermetic, Polarity:Polarized, Tol.:+/-20, Capacitance Va:22.000u, Voltage Rating:15.000D, Pop:387 | 0/9.6061 |
| | | | -P#:CSR13BE225K, Mil#:M39003-01-2283, NSN:5910-00-007-2002, Mfr:Various, Pkg.:Metal, Herm.:Hermetic, Polarity:Polarized, Tol.:+/-10, Capacitance Va:2200000.000p, Voltage Rating:20.000D, Pop:258 | 0/6.4041 |
| | | | -P#:CSR13E156KM, Mil#:M39003-01-2289, NSN:5910-00-932-4455, Mfr:Various, Pkg.:Metal, Herm.:Hermetic, Polarity:Polarized, Tol.:+/-10, Capacitance Va:15.000u, Voltage Rating:20.000D, Pop:645 | 0/16.0102 |
| | | | -P#:CSR13E156KM, Mil#:M39003-01-2290, NSN:5910-00-283-3092, Mfr:Various, Pkg.:Metal, Herm.:Hermetic, Polarity:Polarized, Tol.:+/-20, Capacitance Va:13.000u, Voltage Rating:20.000D, Pop:258 | 0/6.4041 |
| | | | -P#:CSR13E476KM, Mil#:M39003-01-2295, NSN:5910-00-861-5108, Mfr:Various, Pkg.:Metal, Herm.:Hermetic, Polarity:Polarized, Tol.:+/-10, Capacitance Va:47.000u, Voltage Rating:20.000D, Pop:129 | 0/3.2020 |
| | | | -P#:CSR13E476KR, Mil#:M39003-01-2296, NSN:5910-00-113-5689, Mfr:Various, Pkg.:Metal, Herm.:Hermetic, Polarity:Polarized, Tol.:+/-20, Capacitance Va:47.000u, Voltage Rating:20.000D, Pop:129 | 0/3.2020 |
| | | | -P#:CSR13E336KM, Mil#:M39003-01-2304, NSN:5910-00-465-0312, Mfr:Various, Pkg.:Metal, Herm.:Hermetic, Polarity:Polarized, Tol.:+/-10, Capacitance Va:33000000.000p, Voltage Rating:35.000D, Pop:774 | 1/19.2122 |
| | | | -P#:CSR13F476KM, Mil#:M39003-01-2313, NSN:5910-00-460-0843, Mfr:Various, Pkg.:Metal, Herm.:Hermetic, Polarity:Polarized, Tol.:+/-20, Capacitance Va:47.000u, Voltage Rating:35.000D, Pop:387 | 0/9.6061 |
| | | | -P#:CSR13G224KM, Mil#:M39003-01-2344, NSN:5910-00-866-3153, Mfr:Various, Pkg.:Metal, Herm.:Hermetic, Polarity:Polarized, Tol.:+/-10, Capacitance Va:2200000.000p, Voltage Rating:50.000D, Pop:645 | 0/16.0102 |

Reliability Analysis Center (RAC) • 201 Mill St., Rome, NY 13440-6916 • 315-337-0900

AN APPLICATION OF *PROFILER* FOR MODELING
THE DIFFUSION OF ALUMINUM-COPPER ON A SILICON SUBSTRATE

Matthew E. Edwards
Associate Professor Of Physics
Department of Physics

Spelman College
350 Spelman Lane
Atlanta, Ga. 30314-4399

Final Report For:
Summer Research Extension Program
Rome Laboratory

Sponsored By:
Air Force Office of Scientific Research
Bolling AFB, Washington, D. C.

and

Spelman College
Atlanta, Ga. 30314-4399

December 1997

AN APPLICATION OF *PROFILER* FOR MODELING
THE DIFFUSION OF ALUMINUM-COPPER ON A SILICON SUBSTRATE

Matthew E. Edwards
Associate Professor of Physics
Department of Physics
Spelman College

Abstract

Electromigration, the undesirable movement of constituent atoms of the medium, is a deleterious phenomenon for microcircuits and interconnects. In both cases, the electrical currents, of the involved circuits, literally induce and perpetuate the condition of atomic migration. To better understand this phenomenon, and to seek ways to contain or arrest its progression, systematic considerations are needed to measure and predict its occurrence, and methodologies are needed to halt its intrusion. Both theoretical and experimental considerations of interdiffusion, under thermal stressing, are viable activities for addressing electromigration. In regards to theoretical considerations, this research has determined that the fixed-code program, *PROFILER*, is grossly inadequate and inapplicable for describing Auger Spectroscopic data of thin films of Al-Cu on silicon substrates. However, the program is applicable to idealized situations of flat interfaces and constant functional forms. Therefore, initial efforts have been made to develop *PROFILER II*, a more powerful and extensive computer program that would be able to handle variable interfaces, with the inclusion of a data smoothing component through convoluted averaging. To that end, essential issues of non-equilibrium statistical mechanics have been reviewed, studied, and described. The U.S. military protection of its citizens, as provided by the Air Force, will be enhanced by having its electronic devices guarded against electromigration, which can be understood through the anticipated capabilities of the computer code, *PROFILER II*.

AN APPLICATION OF *PROFILER* FOR MODELING THE DIFFUSION OF ALUMINUM-COPPER ON A SILICON SUBSTRATE

Matthew E. Edwards

Introduction

Interdiffusion is an important process in reliability physics. It is the process that is occurring when the components in connecting alloys diffuse across the region of coupling. This process finds itself applicable in issues associated with interconnects, composite materials, high-temperature coatings, and thin-film devices [1-6]. Its importance for the reliability of interconnects is in the prevention or at the very minimum the prediction of electromigration, where the latter is an undesirable phenomenon of atomic migration. The region of atomic migration is referred to as the diffusion layer. This layer is often the region where corrosion, electrical anomalies, embrittlement and other deleterious processes occur. In the cases of high-temperature coatings and thin-film devices, for instance, interdiffusion can lead to early electronic device failures. Interdiffusion, as brought on by electromigration from currents in circuits, can also lead to device failure. Therefore, there is a need to better understand the nature of metallic diffusion, and to have predictive analyses of the diffusants' concentrations as a function of time and penetration depth.

In this investigation, the applicability of *PROFILER* has been extensively studied with the essential outcome that this fixed-code program is inappropriate for describing

variable-interfaced thin films of Al-Cu on a silicon substrate as provided by Auger spectroscopy. Therefore, initial efforts have been made to develop *PROFILER II*, such that the latter will be able to handle variable interfaces, with a component for smoothing the spectroscopic data.

Discussion

Since the application of *PROFILER* to flat interfaces is a valid and a significant operation, the details of its features are presented in the following. Each item is described to varying degrees (See Table 1. for a layout of *PROFILER'S* main menu):

1. Load - This sub-menu item reactivates existing or previously prepared files into the operating program. Pressing <Enter> displays on the screen all file names that have been saved in prior sessions. Select the desired file and press <Enter>.

Note: The Load entry should not be selected until files have been appropriately saved. If selected before files are saved, the program gives an error message.

(The user should start with the New screen, as described later, to avoid this problem).

2. Save - This item saves the current data. The program prompts the user to enter a file name having up to 8 letters or numbers. In all cases, saved information items or selections to *PROFILER* are implemented by pressing <F2>.

3. Save As - This item allows the user to change the name of an existing file that is currently running. The existing file name appears after Save As on the main menu.

Table 1: *PROFILER'S* Main Menu

| <u>Files</u> | <u>Data</u> | <u>*Graphics</u> | <u>*</u> |
|----------------------|---------------------------------|---------------------------------------|----------|
| 1. Load | 7. System Information | 15. **Diffusion Time | 1 |
| 2. Save | 8. ****L matrix | 16. Concentration Differences | 1 |
| 3. Save As | 8.1 Frequency Factors | 17. ***Display Concentration Profiles | 2 |
| 4. New | 8.2 Activation Energies | | 2 |
| 5. Exit (save first) | 8.3 Tracer Diffusivities | | 2 |
| 6. Quit (no save) | 8.4 L matrix | | |
| | 9. ****G matrix | | |
| | 9.1 Regular Solution Parameters | | |
| | 9.2 G matrix | | |
| | 1. D matrix | | |
| | 11. r mEigenvectors of r | | |
| | 12. Alpha matrix | | |

*Selectional only after System Information fill-in sheet and the D matrix have been completely implemented.

**A diffusion time always exist, having a default value of 360,000 seconds.

***Further selective only after concentration Differences have been specified.

****Has second level menu items as shown.

4. New - This item is used to create a new file. It should be the first selection made for a beginning session with **PROFILER**. When selected, a new System Information Screen is displayed and the user must enter the following information about the diffusion couple:

- a. The number of alloying elements (diffusants) in the couple. The program has a default value of 2.
- b. The temperature (in Kelvin) at which diffusion takes place. The program has a default value of 1500K.
- c. Average mole fractions of the solutes in the couple. Note: The program calculates the average mole fraction of the host element, from information about the solutes.
- d. Abbreviations for each alloying element (optional).
- e. MO (Structural Factor) -This item has a default value of zero. It gives information on the diffusion steps which may go backwards rather than forward. Typical values are as follows:

| | | | |
|---------------------|--------|---------------------|---------|
| Simple cubic | - 3.77 | Body Centered Cubic | - 5.33 |
| Face Centered Cubic | - 7.15 | Diamond | - 2.00. |

If the appropriate information is not entered on the System Information Screen, the program gives an error message and automatically stops running.

5. Exit (save first) - This item will exit the program and save the data under the user's specified filename.

6. Quit (no save) - This item will exit the program without saving the entered data. The program has a safety feature which prompts the user to save the data before exiting the program.

7. System Information - This item is the same as that available under # 4 above. It gives pertinent information about the current diffusion couple, and can be selected at any time while using the program.

8. L matrix - This item is where tracer diffusion data are entered.

8.1 Frequency factor (A) - (Not considered in this investigation. The program runs and gives satisfactory results without specifying the frequency factor.)

8.2 Activation energies (Q) - The typical energies for migration of involved atoms

8.3 Tracer diffusivities (D) - The typical diffusion coefficients of involved atoms

8.4 L matrix - (Not considered in this investigation. The program runs and gives satisfactory results without specifying the L matrix.)

9. G matrix - This is a pull-down menu to enter thermodynamic information about the diffusants or atoms.

9.1 Regular Solution Parameters - These values can be approximately related to heats of mixing, ΔH^{mix} , mole fractions, N_i , and other measured thermodynamic quantities by various formulas, e.g.:

$$\omega_{ij} \cong \frac{\Delta H^{mix}}{N_i N_j}.$$

9.2 G matrix - The second derivative of the free energy are entered here in units of J/mole, where

$$G_{ij} = \frac{\partial^2 G}{\partial N_i \partial N_j}.$$

10. D matrix - The diffusivity matrix elements [6-8] are entered here if they are known.

11. r matrix - The square root diffusivity matrix [r] is entered here.

Note: If L and G matrices are provided, then **PROFILER** will automatically calculate D and r. If D or r is provided, the program will calculate the one that's not entered. The calculated diffusivities appear on their respective menu screens.

12. Eigenvalues of r - The square roots of the eigenvalues of D are displayed

(i.e. $r_i = \sqrt{D_i}$).

13. Eigenvectors of r - The α^{-1} matrix is displayed. Columns of this matrix are

eigenvectors of both D and r. The α and α^{-1} matrices diagonalize D by the transformation, $D_i = \alpha D \alpha^{-1}$.

14. Alpha matrix - This item displays the α matrix.

15. Diffusion Time - The isothermal heat treatment time is entered in units of seconds. The program has a default time of 360,000 seconds (100 hours).

16. Concentration Differences - The initial concentration differences between the diffusing couple are entered. The values are obtained by subtracting concentrations on the left side of the coupling interface from those on the right side.

17. Displays Concentration Profile - The monitor's screen shows the concentration profiles of each solute (diffusant). The axes variable are the followings:

X axis - Gives the distance from the initial interface (a distance of 10 μm is between graduations marks on the X axis).

Y axis - Gives the difference in concentration between the local concentration and the average concentration for each solute. (0.2 of units of the concentration are between graduations on the Y axis).

The purpose of the Numeric menu is to create a file of concentration profiles in a tabular form. This file can then be imported into a spread sheet or plotting program to obtain a hard copy [9] of the concentration profiles or to construct diffusion paths.

18. Diffusion Time - Displays the time as entered in # 15 above.

19. Concentration Differences - Displays the initial concentration differences as used to calculate the concentration profiles.

20. View matrix A - The concentration profile for an n component diffusion couple is given by the sum

$$C_i(x,t) = C_i^R - \sum_{j=1}^{n-1} A_{ij} \operatorname{erfc}\left(x / (2\sqrt{D_j t})\right),$$

where C_i^R is the initial concentration of solute i in the alloy on the right side of the initial interface, and D_j is one of the n-1 eigenvalues of the D matrix [5-8]. The D matrix gives

the presentation of the A_{ij} coefficients of the complementary error function, $\operatorname{erfc}\left(\frac{x}{2\sqrt{D_j t}}\right)$.

21. Generate Numeric Data -- Commas - This item creates a file with numeric data separated by commas.

22. Generate Numeric Data -- Spaces - This item creates a file with numeric data separated by spaces.

The format of the output file is the following:

| column 1 | column 2 | column 3 | column n. |
|----------|----------|----------------|-----------|
| -0.00100 | 0.2 | 0.3 | -1.4 |
| ... | ... | ... | ... |
| ... | ... | ... | ... |

Column 1 gives x axis values in centimeters (x = 0 is the original-finite interface between the alloying couple).

Column 2 gives concentration differences for the first solute between the local concentration and the average initial concentration , where

$$\Delta C_i = C_i(x,t) - \frac{C_i^R + C_i^L}{2} .$$

(The concentration difference is always zero at x = 0).

Column 3 gives the concentration differences for the second solute.

Column n gives the concentration differences for the n-1 solute.

Using the above menu items, **PROFILER** has been studied to asset its ability to depict the non-equilibrium movement of atomic particles (Aluminum and Copper Atoms) on a Silicon substrate. The program functions adequately for flat interfaces. However, real coupling at interfaces, as represented by Auger Spectroscopic data, is such that it is rarely if ever completely flat. Thus the question becomes - would it be economically feasible to write a differ computer code for the varied interfacial configurations? We supplied a yes to

this question, and set-about initiating the development of such a program. To that end, we reviewed non-equilibrium statistical mechanics, considering such issues of time correlation functions, relaxation to equilibrium and stochastic processes.

The following table gives essential information on Al and Cu that has been useful in the development of the extended computer program. It gives data on coordination numbers, lattice parameters, etc.

| Name | Density (g/cm ³) | Atomic Number | Crystal Structure | Lattice Para- meter(A) | Distance of Closest Appro- ach | Melting Point(C) | Coordin- ation* |
|------|---------------------------------|------------------|----------------------|------------------------------|--|----------------------|--------------------|
| Al | 2.70 | 13 | FCC, Al | 4.0490 | 2.862 | 660 | 12 |
| Cu | 8.90 | 29 | FCC, Al | 3.6153 | 2.556 | 1083 | 12 |

Table 2. Essential Properties of Al and Cu. *Coordination number is the number of nearest neighbor atoms.

Methodology

The outcome from **PROFILER** may be obtained from one of two different modes of operation, depending on the type and amount of information that's initially known about the diffusing couples [9-20]. Table 3, below, outlines the two different modes of operation. **PROFILER II** should be developed with similar modes of operation.

| USING PROFILER TABLE 3 | |
|-----------------------------------|---------------------------------|
| Method I. | Method II |
| Pre-Determined Diffusivity Method | Undetermined Diffusivity Method |
| 1. System Information | |
| a. Number of Components in the | |

interdiffusion-couple. n has the value: $(1 < n < 8)$, and must be the first entry.

(Same as Method I for a - d.)

b. Temperature in Kelvin at which the diffusion process occurs.

c. Average Mole fraction (at. Pct.) of each solute between the left member and the right member of the interdiffusion couple. (At this point the program calculates the average fraction of the solvent, the "host" element).

d. Enters abbreviations of alloying element. (optional entry - but helpful for accountability)

e. Structure Factor (MO) Unnecessary Entry for this method. (The appropriate value has already been considered in the Pre-determination of the diffusivity)

e. Structure Factor (MO) As determined from diffusion theory:

1. simple cubic - 3.77
2. body centered cubic - 5.33
3. Face centered cubic - 7.15
4. Diamond - 2.00

2. Diffusivities Not entered.

2. a. Tracer diffusion Values (Highlight L matrix on menu bar). Pre-exponential (A) units of cm^2/sec , Activation Energies (Q) units of kcal/mole

b. Thermodynamic Data - Heat of mixing ΔH^{mix} , Mole fraction of N_I and N_j of solutes

3. Diffusion Time in seconds

3. (Same as method I)

4. Concentration Differences

4. (Same as method I)

Enter the concentration of right member of the interdiffusion couple minus the concentration of the left member.

5. Displays Concentration differences

5. (Same as method I.)

(difference between the absolute concentration minus the average as entered for the diffusion couple):

Ordinate axis, in units of at. Percent

Abscissa axis, in units of cm.

With the above procedure understood on just how to use *PROFILER*, our efforts became one of developing directly and/or obtaining from the scientific literature a scheme for handling the variation in the coupling region and the overall lack of a step function behavior for the diffusants of the system, copper and aluminum on a silicon substrate. A scheme to reduce the random fluctuations in the Auger Spectra was obtained and used[21-25]. This scheme used the method of convoluted averaging where the convolution function was obtained from a least squares procedure. This resulted in an invaluable smoothing of the data but not in an overall reduction in the interfacial variation and/or the lack of a constant function behavior for the concentrations away from the interface. Fig. 1 depicts the dilemma of the problem showing the typical nature of the Auger Data. The concentration function for Al is fairly constant from a depth of zero nanometers (nm) up to 900 nm. Then the function varies decreasingly in the cross-over region (a distance of ≈ 200 nm). This distance is more than 600 times the distance of closest approach in the aluminum structure.

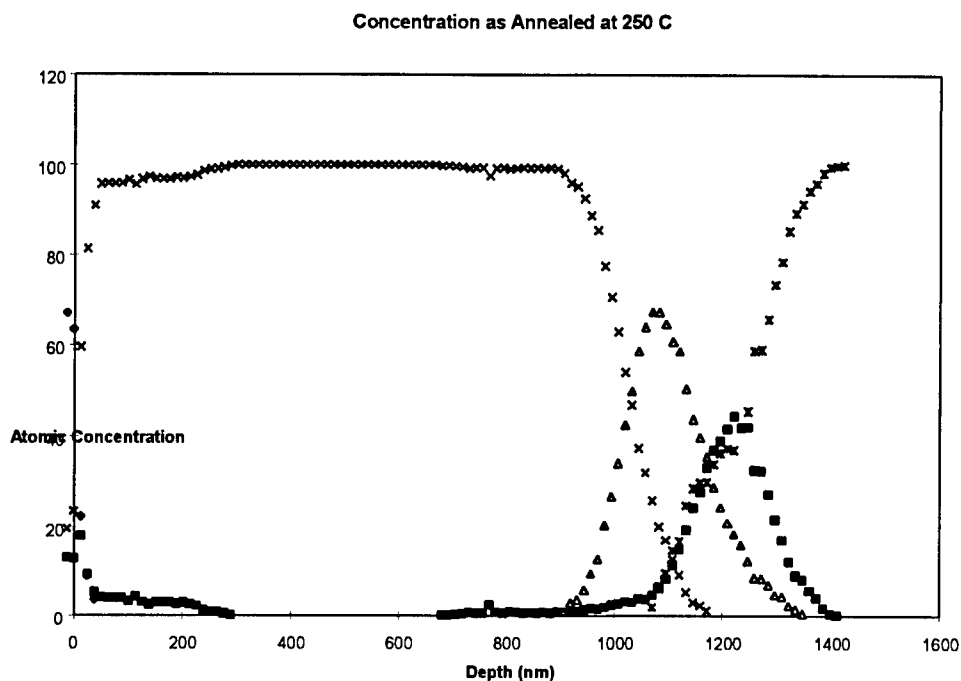


Fig. 1. Shows the Concentration Profiles as a Function of Depth into the Sample.

This yields a transitional interface of some 200 nm, as opposed to an abrupt situation. For such a thin film, this variational region is too large to be considered as being insignificant in size. Also, Fig. 1 shows that the presence of copper sets-in at a depth of about 900 nm and continues over through 1350 nm, in the shape of a peaking function at about 1100 nm. The presence of silicon begins to show up at a sputtering depth of 700 nm, out to and beyond 1400 nm. Oxygen and carbon are minimally present as impurities. Fig. 2 is a reproduction of Fig. 1 with the sputter time (at a rate 6.3 nm/min) as the abscissa.

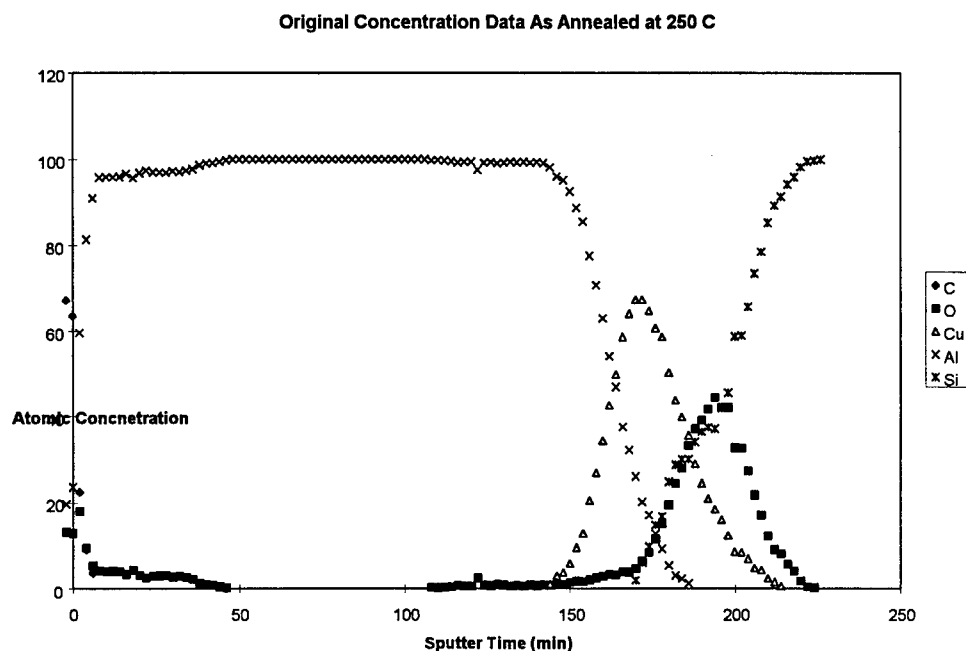


Fig. 2. Shows the Concentration Profiles As Produced By the Auger Spectrometer

From the typical Auger Spectra, it became self-evident that our efforts could be better served by initiating the procedure for developing *PROFILER II*, a computer program having greater versatility in general interfacial regions and allowing for non-constancy away from the interfaces.

To that end, the literature was explored in non-equilibrium statistics (transport theory), equilibrium statistics and thermodynamics for procedures of obtaining and applying the diffusion equation to various interfacial situations. The literature [26-41] was discovered to be replete with information on the diffusion equation as it is used in simple geometrical situations. But not much details were observed on comparative

derivations. Atlas! We uncover the idea that diffusion is just another depiction of the random - walk problem and, as such, the diffusion equation can be derived from this point of view, with the continuation of solving the same in very general geometrical situations.

Therefore, we have obtained the derivation of the diffusion equation in one dimension in the following. The derivation is succinct; yet complete. It is presented below.

The non-equilibrium process of diffusion, in a liquid or solid, as in a gas, may be calculated from a random - walk description. However, the physical model is somewhat different for a liquid or a solid. In a solid, in particular, a diffusing particle is at an equilibrium position with respect to the forces exerted by neighboring particles. Nearby, to the site in question, there are m other similar positions or sites, separated from the given position by a potential barrier, which determines the probability per unit time that the particle jumps to one of these available sites. Let λ_i be the distance of the i th new position from the given equilibrium position, let p_i be the probability that a particular jump is to the i th new position, and let θ_i be the angle between the x - axis and the vector from the given position to the i th position. The average distance a particle moves in the x direction as a result of the k th jump is

$$d_k = \langle x \rangle = \left(\sum_{i=1}^m p_i \lambda_i \cos \theta_i \right)_k \equiv \langle \lambda \cos \theta \rangle$$

where the fences $\langle \rangle$ denote ensemble averaging. This d_k value is zero if there is no external field to make the probabilities differ for jumps in the positive or negative x - directions. The average distance moved in the x direction after n jumps is n times $\langle \lambda \cos \theta \rangle$, which also vanishes. On the other hand, the average square of the distance moved in the x-direction is

$$\langle x^2 \rangle = \sum_{h,k} d_h d_k = \sum_h d_h^2 = n \langle \lambda^2 \cos^2 \theta \rangle$$

This equation is valid since there is no correlation between distance moved on different jumps, that is, $\langle d_h d_k \rangle = 0$ for $h \neq k$. Furthermore, if there is no correlation between λ_i and $\cos \theta_i$, which is assumed to be the case in this derivation, then

$$\langle x^2 \rangle = n \langle \lambda^2 \rangle \langle \cos^2 \theta \rangle$$

It is well known that the average value of $\cos^2 \theta$ is 1/3, and the average value of λ^2 is λ^2 itself (the average of a constant is that constant). Thus the random - walk model leads to the result that the average square distance moved by a particle is proportional to the number of steps taken and to the square of the average step length. This condition on the average square distance is valid if the diffusion coefficient $D = \frac{1}{2} s \lambda^2$ where s is the number of steps taken per unit time. If we combining Fick's first law, that the flux of

particles is equal to $-D$ multiplied by the concentration gradient, with the equation of continuity, we obtain Fick's second law (the diffusion equation in one dimension),

$$\frac{\partial \rho}{\partial t} = \frac{\partial}{\partial x} \left(D \frac{\partial \rho}{\partial x} \right)$$

Here, $\rho = \rho(x, t)$ is the number density or concentration of particles, $\rho(x, t)dx$ being number of particles between x and $x+dx$ at time t . It is normalized according to

$$\int_{-\infty}^{+\infty} dx \rho(x, t) = N$$

where N is the total number of particles. Note that $\rho = N$ times the Probability density function. If D can be assumed independent of x , Fick's second law simplifies to

$$\frac{\partial \rho}{\partial t} = D \frac{\partial^2 \rho}{\partial x^2}$$

Normally, D depends on concentration, and diffusion occurs in the presence of a concentration gradient, so D will depend on x . If the concentration is low, this dependence may be unimportant. In solids and alloys, involving tracer diffusion, D may be taken to be constant since the tracer particles are the same as the other particles except for labeling.

In preparing for the development of *PROFILER II*, we considered the solutions to Fick's second law, which depend on initial conditions, in two simple geometrical situations, the

point function and the step function. If $\rho = N\delta(x)$ at $t=0$ (all particles are at $x=0$ to start, the point function), the solution is

$$\rho = N(4\pi Dt)^{-1/2} \exp\left[\frac{-x^2}{4Dt}\right]$$

where $\langle x^2 \rangle = 2Dt$. On the other hand the step function has the initial conditions

$$\rho = \rho_0 \text{ for } x \leq 0 \text{ and } \rho = 0 \text{ for } x > 0 ,$$

(particles are on one side of the finite plane at $x=0$). The solution can be found by considering this initial density profile as a superposition of δ -function profiles, so that for $t>0$ one can just superpose density profiles of the form as given above. If the step-function profile is written as

$$\rho(x,0) = \rho_0 \int_{-\infty}^0 \delta(x-y) dy$$

the density at $t>0$ is

$$\rho(x,t) = \rho_0 \int_{-\infty}^0 (4\pi Dt)^{-1/2} \exp\left[\frac{-(x-y)^2}{4Dt}\right] dy$$

On substituting $z = x - y$ as the variable of integration, this becomes

$$\begin{aligned}
\rho(x,t) / \rho_0 &= (4\pi Dt)^{-1/2} \int_{-\infty}^0 \exp\left[\frac{-z^2}{4Dt}\right] dz \\
&= \frac{1}{\sqrt{\pi}} \int_{x/2\sqrt{Dt}}^{\infty} e^{-y^2} dy = \frac{1}{2} \frac{2}{\sqrt{\pi}} \int_{x/2\sqrt{Dt}}^{\infty} e^{-y^2} dy \\
&= \frac{1}{2} \operatorname{erfc}\left(\frac{x}{2\sqrt{Dt}}\right) = \frac{1}{2} \left[1 - \operatorname{erf}\left(\frac{x}{2\sqrt{Dt}}\right) \right]
\end{aligned}$$

where $\operatorname{erf}\left(\frac{x}{2\sqrt{Dt}}\right)$ is the error function. Mathematically, this function is defined to be as

$2 / \sqrt{\pi}$ times the integral from 0 to $\frac{x}{2\sqrt{Dt}}$ of $\exp(-y^2)$ so that $\operatorname{erf}(\infty) = 1$. The above

expression for $\rho(x,t) / \rho_0$ is plotted in the results section, for $\sqrt{Dt} = 0.1, 0.5, 0.9, 1.3, 1.7$

where the appropriate series expansions are used for the error function. These procedures

have provided a beginning for the extension of **PROFILER** to a new and more powerful

computer program, **PROFILER II**.

Results

A successful smoothing operation was obtained from the equation

$$Y_j^* = \frac{\sum_{i=-m}^{i=m} C_i Y_{j+i}}{N}$$

where m is the number of terms used in the convolution, and C_i 's are the terms of the convolution from the method of least squares. This procedure, however, did not remove the difficulty of the variable interface of the coupling diffusants. Therefore, we initiated the development of a self-contained computer program with appropriate capabilities that are lacking in *PROFILER*. Specifically, we obtained the diffusion equation from the general principle of random - walk, and obtained its solution in two important situations: the point function and the step function.

Using the series expansions for the error function, we obtained the graphs of concentration, as depicted in fig. 3 for five cases of \sqrt{Dt} . These graphs show the manners in which diffusion takes place in the step function situation. Note that as

\sqrt{Dt} gets larger, greater movement of the

atoms occurs for given values of x . The mathematical representations of the erf function

as used in Fig. 3. are:

$$\text{erf}(a) = \frac{2}{\sqrt{\pi}} \left(a - \frac{a^3}{3 \cdot 1!} + \frac{a^5}{5 \cdot 2!} - \frac{a^7}{7 \cdot 3!} + \frac{a^9}{9 \cdot 4!} - \frac{a^{11}}{11 \cdot 5!} + \frac{a^{13}}{13 \cdot 6!} \right) \quad \text{valid for } a \leq 1$$

and

Step Function Diffusion

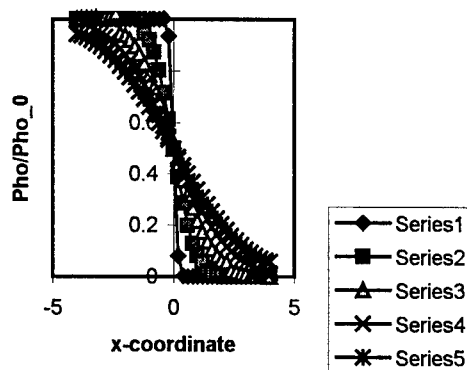


Fig 3. Solution to diffusion equation for various values of \sqrt{Dt} , starting with a step profile.

$$\operatorname{erf}(a) \approx 1 - \frac{e^{-a^2}}{\sqrt{\pi}a} \left(1 - \frac{1}{2a^2} + \frac{1 \cdot 3}{(2a^2)^2} - \frac{1 \cdot 3 \cdot 5}{(2a^2)^3} + \frac{1 \cdot 3 \cdot 5 \cdot 7}{(2a^2)^4} - \frac{1 \cdot 3 \cdot 5 \cdot 7 \cdot 9}{(2a^2)^5} \right) \text{ valid for } a \geq 1$$

the asymptote region

Continuation in the development of **PROFILER II** to handle variable initial interfaces, even statistically fluctuating interfaces, is an obtainable and desirable goal.

Conclusion

PROFILER, as a fixed-code computer program, has been shown to be a powerful and efficient tool for modeling the diffusion of metallic atoms across a flat initial plane of coupling for two alloying components. It produces the concentration profiles in a matter of a few seconds for up to eight diffusants in the alloying couple. It is more than adequate for such situations. However, its usage for variable interfaces is inappropriate and ineffective. Also, when using **PROFILER**, data can be convolutedly averaged before reduction to lessen statistical noise, without degrading the Auger signal.

Faced with the dilemma of the fixed-code of **PROFILER**, we initiated the process of developing **PROFILER II**, and have progressed to the point of developing the analytical functions for the point source and the step functions. These features form the basis for the actual development of a new computer program for the trace diffusion of atoms of thin films on substrates.. It remains a desirable task to complete the development of such a computer program.

References

1. **PROFILER: Diffusion Couple Software to Predict Concentration Profiles and [D]**, as written by William B. Brockman, at the University of Connecticut, and Available from John E. Morral, Dept. of Metallurgy and Institute off Material Science, University of Connecticut, tel: 203 486-2923, FAX 203 486-4745
2. *Elements of Material Science*, 2 nd. Ed. Lawrence H. Van Vlack, Addison-Wesley Publishing Co., 1964.
3. *Copper Diffusion Into Aluminum-Silicon Metallization By Accelerated Thermal and Electrical Stressing*, G.O. Ramseyer, L.H. Walsh, J.V. Beasock, H.F. Helbig, R.C. Lacoe, and S. Brown, in press.
4. *Handbook of Auger Electron Spectroscopy*, 2 nd. Ed., Lawrence E. Davis, et. Al., Physical Electronics Division, Perkin-Elmer Corporation, 1976.
5. *Classification Of Concentration Profiles In Quaternary Diffusion Couples*, M.K. Stalker, and J.E. Morral, Acta Metall. Mater. Vol 38, No. 3, pp 439-447, 1990.
6. *Applications of the Square Root Diffusivity to Diffusion in Ni-Al-Cr Alloys*, M.S. Thompson, J.E. Morral, and A.D. Romig, Jr., Metallurgical Transactions A, Vol 21A, Oct. 1990, p 2679.
7. *MULTICOMPONENT DIFFUSION: Implementation of the Square-Root Diffusivity Method Via the PROFILER Computer Program*. T.H. Cohen and M.E. Glicksman, Modeling Simul. Mater. Sci. Eng. **3**, (1995) 585-596.
8. *Interdiffusion in the Ni-Cr-Co-Mo System at 1300 C*, J.A. Heaney, III and M.A. Dayananda, Metallurgical Transactions A, Vol 17A, June 1986, pp 983
9. Microsoft Manuals on Microsoft Windows95, Word 7.0, Excel 7.0, Equation Editor, etc.
10. Private Communications with Professor Morral, University of Connecticut
11. *The Analysis of PROFILER for Modeling the Diffusion of Aluminum-Copper on a Silicon Substrate*, Matthew E. Edwards, Air Force Office of Scientific Research, In Press.
12. *Elements of X-Ray Diffraction*, B.D. Cullity, Addison-Wesley Publishing Co, 1978.

13. Characterization of Copper Diffusion Into Al and Al-1% Si Polycrystalline Thin Films, L.H. Walsh, G.O. Ramseyer, J.V. Beasock, H.F. Helbig, In Press.
14. *Handbook Of Chemistry and Physics*, 6 th. Ed., CRC Press, 1985-86.
15. *Advanced Mathematics For Engineers and Scientists*, Murray R. Spiegel, Schaum's Outline Series, McGraw-Hill, 1993.
16. *Phase Transformation in Metals and Alloys*, D.A. Porter and K.E. Eastering, Van Nostrand Reinhold, 1982.
17. Butkov, Eugene *Mathematical Physics* (Addison-Wesley) Publishing Co., 1968.
18. Skoog, D and West. D., *Principles of Instrumental Analysis* (Sounders Golden Sunburst Series, Philadelphia, 1980.
19. Arken, G., *Mathematical Methods for Physicists* (Academic Press, Inc., 1985)
20. Dautray, R. and Lions, J.-L., *Mathematical Analysis and Numerical Methods for Science and Technology* (vol. 6 Evolution Problems II The Navier-Stokes and Transport Equation and Numerical Method and vol. 4, Integral Equations and Numerical Methods) Springer-Verlag, New York, (1992).
21. Press, W., et al., *Numerical Recipes: The Art of Scientific Computing*, (Cambridge Press, New York, 1991).
22. Bartee, T., *Basic Computer Programming*, Second Edition, (New York, Harper & Row, Publishers, 1985)
23. Spiegel, Murray, *Mathematical Handbook - of Formulas and tables-* (New York, Schaum's Outline Series, McGraw-Hill book Co, 1968)
24. Bevington, Phillip R., *Data Reduction and Error Analysis for the Physical Sciences* (New York, McGraw-Hill Book Co., 1969)
25. Savitzky, Abraham and Marcel J.E. Golay, "Smoothing and Differentiation of Data by Simplified Least Squares Procedures", *Analytical Chemistry*, Vol. 36, No. 8, July 1964.
26. Rayne, J. A. and M. P. Shearer, and L. Bauer, "Investigation of interfacial reactions in thin film couples of Aluminum and Copper by measurement of low temperature

contact resistance", *Thin Solid Films*, **65**, (1980), 381-391.

27. Goodisman, Jerry, *Statistical Mechanics for Chemists* (New York, John Wiley & Son, Inc., 1997)
28. Reif, Frederick, *Statistical and Thermal Physics, International Edition* (New York, McGraw-Hill Book Co., 1985)
29. Reif, Frederick, *Statistical Physics (Berkeley Physics Course, Vol. 5, McGraw-Hill, 1965)*
30. Koonin, Steven, *Computational Physics* (New York, Addison-Wesley Co., 1992)
31. Espinola, Thomas, *Introduction to Thermophysics* (Dubuque, Wm. C. Brown, Publishers, 1994)
32. Kittel, Charles and Herbert Kroemer, *Thermal Physics, 2 nd. Edition* (New York, W. H. Freeman and Co., 1980)
33. Daniel, Cuthbert and Fred S. Wood, *Fitting Equations To Data - Computer Analysis of Multifactor Data-* (New York, John Wesley & Sons, 1980)
34. Microns, Ronald E., *Difference Equations* (New York, Van Nostrand Reinhold Co., 1987)
35. Dwyer, V. M., F.-S. Wang, and P. Donaldson, "Electromigration Failure in a Finite Conductor With a Single Blocking Boundary", *J. Appl. Phys.* **76**, (11), December, 1994, page 7305
36. Edwards, M. E., Y.H. Hwang, and X.-L. Wu, "Large Deviation From the Clausius-Mossotti in a Model Microemulsion", *Physical Review E*, **49**, No. 5, 1994, page 4263
37. P.G. De Gennes, and C. Taupin, *Microemulsions and the Flexibility of Oil/Water Interfaces*", *The Journal of Physical Chemistry*, **86**, No. 13, 1982, page 2294
38. Edwards, M. E., X. L. Wu, J. S. Wu, J. S. Huang, and H. Kellay, "Electric Field Effects on a droplet Microemulsion", *Physical Review E*, **57**, No. 1, 1998, page 799.
39. Saffman, P.G., "Brownian Motion In Thin Sheets of Viscous Fluid", *J. Fluid Mech.* **73**, part 4, 1976, page 593

40. Rothman, S., and N. L. Peterson, *Diffusion In Dilute Alloys*, Metals Handbook, American Society of Metals, 1985.
41. Clark, Noel, J. H. Lunacek, and G. B. Benedek, "A Study of Brownian Motion Using Light Scattering" Amer.J. Phys. **38**, No. 5, 1970, page 575

ANALYSIS OF STRESSED SPEECH USING CEPSTRAL DOMAIN FEATURES

K. Gopalan
Professor
Department of Engineering

Purdue University Calumet
Hammond, IN 46323

Final Report for:
Summer Research Extension Program
Rome Laboratory

Sponsored by:
Air Force Office of Scientific Research
Bolling Air Force Base, DC

Rome Laboratory

and

Purdue University Calumet

December 1997

Analysis of Stressed Speech using Cepstral Domain Features

K. Gopalan
Professor
Department of Engineering
Purdue University Calumet
Hammond, IN 46323

Abstract

The effect of workplace stress on speech parameters was studied in this project. Correlation between certain cepstral coefficients and the heart rate of fighter aircraft flight controllers was investigated. It was found that as the heart rate increased due to stress, mel cepstral coefficients corresponding to pitch range of frequencies showed increasing periodicities in their correlated values. In addition, the index of peak coefficients increased with increasing heart rate while the standard deviation of cepstra at all the indices, in general, showed a decreasing trend with increasing heart rate. As with pitch frequencies, cepstral indices at which proportional variation with heart rate occurred, depended on the speaker and the phoneme.

Analysis of Stressed Speech using Cepstral Domain Features

K. Gopalan

I. Introduction

Analysis of human stress based on the speech signal output of a speaker is important in many applications. These applications include automatic monitoring of the stress levels of callers for emergency services so as to respond with appropriate service personnel, and analyzing the physiological and emotional state of fighter aircraft pilots to decide on their suitability to proceed with their assigned task. In addition, acoustic correlates of stress alter speech- and speaker-dependent parameters; this alteration, unless compensated for in the feature domain, can result in poor performance of speech and speaker recognition systems.

This report presents the results of analyzing certain cepstral domain features as a function of the heart rate of fighter aircraft pilots.

II. Review of Speech Features Affected by Stress

One of the predominant effects of stress on speech signal is the variation in the fundamental frequency, F0. This was demonstrated by Lieberman and Michaels [1] by presenting to a group of listeners only the pitch information available in tapes containing speech at different emotional states. Several researchers have recently found that F0 variations correlated well with stress. In one of the early works by Williams and Stevens [2], the median value and the range of F0 were found to rise for speech uttered in anger or fear from those corresponding to normal speech; for sorrow, the value and the range of F0 were reduced from their neutral values. Using cepstral analysis of pitch, Levin and Lord [3] showed that an increase in the range averaged pitch frequencies indicated emotional change from normal state. Streeter, et al [4] found that a listener's rating of a speaker's stress level correlated positively with average pitch; however, based on their pitch and amplitude measurements, they concluded that no single acoustic pattern resulted from increased stress. Using the Teager energy profile, the change in the nonlinear component of speech was shown by Cairns and Hansen [5] to distinguish clearly between loud and angry speaking styles. While the modulation pattern of the first formant, as quantified by the Teager energy profile, changed appreciably between loud and angry speech, clear speech could not be easily differentiated from neutral speech. Based on listeners' perception of stress in F0-altered speech, Protopapas and Lieberman [6] found that mean and maximum F0 within an utterance correlated significantly with higher stress ratings. They concluded that except for extreme stress and terror, their results could not be generalized. Hansen and Womack [7] used a set of mel-distributed cepstral features and a neural network to classify 11 stress conditions including angry speech, question, soft speech and slow speech. They showed that different stress conditions and phoneme classes affect the features

differently; vowels, in particular, are significantly affected with a high degree of separation using autocorrelation of the cepstral parameters for the different stress conditions.

Another area related to the analysis of stressed speech is concerned with the task of identifying the acoustic-phonetic differences between normal and abnormal speech. Identification of features that distinguish between normal, loud and Lombard speech, for example, is important in the development of robust speech and speaker recognition systems. The recognition systems are, in general, trained using mostly normal speech because of (a) the difficulty of collecting data under abnormal conditions, and (b) the volume of such data. When presented with data collected in a stressful and noisy environment, however, the performance of these systems becomes degraded. Rajasekaran, et al [8] report an order of magnitude decrease in the recognition of speech from a vocabulary of words under simulated stress and noise conditions. Since a large variation in spectral tilt is commonly observed for speech under stress, Chen [9] proposed a cepstral domain compensation to normally trained word models for reducing the recognition error rate. Stressed speech was quantified by smoothing and fitting the means of certain cepstral components to an exponential function. By subtracting the smoothed values from test vectors, significantly low recognition error rates were achieved for a data base of simulated stressed speech. Using a data base of normal, loud (simulated high stress) and Lombard (pink noise injected) speech, Stanton, et al [10] observed energy migrations in different frequency bands and spectral tilt between normal and abnormal speech. By accounting for these variations, Stanton, et al [11] report an improvement in the performance in the recognition of stressed speech. Bou-Ghazale and Hansen [12] took the approach of modeling stressed speech as a sequence of articulator movements whose paths deviated from those for normal speech. Using the deviations to train and test, they report an improvement in the recognition of keywords in simulated stressed speech in noise-free conditions.

From the brief discussion above, it is clear that different conditions of physiological and emotional stress affect speech features differently. Although the fundamental frequency F_0 is generally observed to vary with stress, no single stress token appears to correlate all phonemes under various speaking styles. The difficulty of obtaining actual stressed speech data for training speech recognition systems has led to simulated data that were restricted mostly to loud and Lombard styles. Statistical variations of features in the cepstral domain, in general, have been found to represent well the stress-induced variations for classifying simulated stressed speech.

III. Problem and Approach

In this study, the problem of correlating a measured stress factor with a cepstral domain feature was addressed. The Speech Under Stress database used for this study consists of speech from European (non-English native) fighter controllers under stress. The database contains speech utterances from nine male controllers that were recorded during communication with fighters. Various stress factors, such as systolic pressure, diastolic pressure, carbon dioxide and heart rate, were measured during the speech. For the present study, heart rates were chosen to

relate with speech features because of the nonavailability of other measurements for all the speech files available in the database. Each speech file was encoded in 16 bits per sample at a rate of 16000 samples per second.

Since the database has continuous speech with a large vocabulary of words, it was necessary to choose an utterance that occurred frequently and at varying heart rates for each speaker. Three different heart rates with as large a variation as possible among them were chosen for each speaker. It was found that the utterance "*bull's eye*" most commonly occurred with three different heart rates for many of the speakers.

Using Entropic Waves+ package, the utterance "*bull's eye*" corresponding to each of three different heart rates was extracted for each speaker. Cepstral domain features were computed for the segmented utterances and analyzed with the known heart rates.

IV. Low Quefrency Cepstral Features

Inasmuch as the fundamental frequency F0 is commonly observed to rise in stressed speech, cepstral features corresponding to the range of F0 were studied for each speaker. Table I lists the range of F0 values for the speakers considered¹.

Table I
Range of F0 for the utterance "*bull's eye*"

| Speaker | Minimum F0, Hz | Maximum F0, Hz |
|---------|-------------------|-------------------|
| JO | 142 | 261 |
| RD | 101 | 273 |
| RK | 96 | 180 |
| SK | 90 | 180 |
| VL | 100 | 155 |

Based on the above range of frequencies for F0, mel-cepstra were calculated at the center frequencies of 70 Hz to 300 Hz in steps of 10 Hz with a fixed bandwidth of approximately 19.5 Hz each. A frame length of 15 ms with an overlap of 10 ms was used in segmenting each utterance. After testing with Hanning, triangular and exponentially decaying windows, the final analysis was carried out using Hamming windows. With 4096-point discrete-Fourier transform (giving a frequency resolution of 3.9 Hz), cepstra at 24 points (quefrequencies) over the pitch range of frequencies were calculated at cepstral indices from 18 to 77 in steps of 2 or 3. A triangular bandwidth of 5 points (≈ 19.5 Hz) was used at each index. In addition, another set of cepstra at 30 points corresponding to formant range of frequencies (at approximately 250 Hz to 5000 Hz) were calculated as discussed in Section V. The features

¹ The pitch frequencies were obtained using Waves+ software package from Entropic [15] with the same frame length of 15 ms and overlap of 10 ms as for the cepstral calculations.

formed from the 54 cepstral coefficients for each speaker under three different heart rates were studied for correlation with the heart rates.

Fig. 1 shows the distribution of cepstral coefficients with frames for speaker VL. As given in Table II, the pitch frequencies for the low, medium and high heart rates for this speaker lie within the range of 100 Hz to 155 Hz. This range is the smallest for all the speakers considered. Because of this low range of fundamental frequency variation, only cepstra at indices 26, 28, 31, 33, 36, 38 and 41 (corresponding to frequencies 101.6 Hz, 109.4 Hz, 121.1 Hz, 128.9 Hz, 140.6 Hz, 148.4 Hz, and 160.2 Hz, or straight indices 4 through 10) are shown in Fig. 1. These cepstra were scaled by the total of the logged spectrum over the entire utterance, "*bull's eye*" and normalized to zero mean and unity variance.

For comparison, the scaled mean of each of the 54 cepstral values across all the frames, the total logged spectrum and the standard deviation of the scaled cepstra are shown for the three utterances of the speaker VL in Fig. 2.

Table II
Range of F0 at Measured Heart rates for speaker VL

| Index | Pitch Range, Hz | Heart Rate | Norm. Heart Rate |
|--------|--------------------|------------|---------------------|
| VL065G | 100 - 152 | 103.3 | 1.490 |
| VL045G | 105 - 117 | 100.3 | 1.019 |
| VL093G | 110 - 140 | 97.45 | 0.673 |

It must be noted that the average value of each of the 54 mel cepstral coefficients across all the frames for a speaker is almost constant regardless of the heart rate. Fig. 3 shows the mean and the standard deviation of the cepstral coefficients in the expanded scale for the pitch range of frequencies for speaker VL. For comparison, Fig. 4 shows the mean and the standard deviation of the cepstral coefficients for speaker RD at three different heart rates. The heart rates used and the pitch frequencies at these rates for RD are given in Table III. Note that RD has the widest range of pitch frequency variation; the range of heart rate, however, is one of the smallest. The mean values of each of the cepstral coefficients for this speaker, however, show an increase with the heart rate. From these two cepstral distributions (and from the others not shown, but calculated), it appears that the mean of the scaled cepstra is not a good indicator of stress. It may be inferred that the logged spectral energy, while increasing in frames under stressed conditions, may be decreasing in other frames. Depending on the phonemes (and hence the number of frames involved), therefore, the overall mean of each cepstral value across an entire utterance may vary nonuniformly with heart rate.

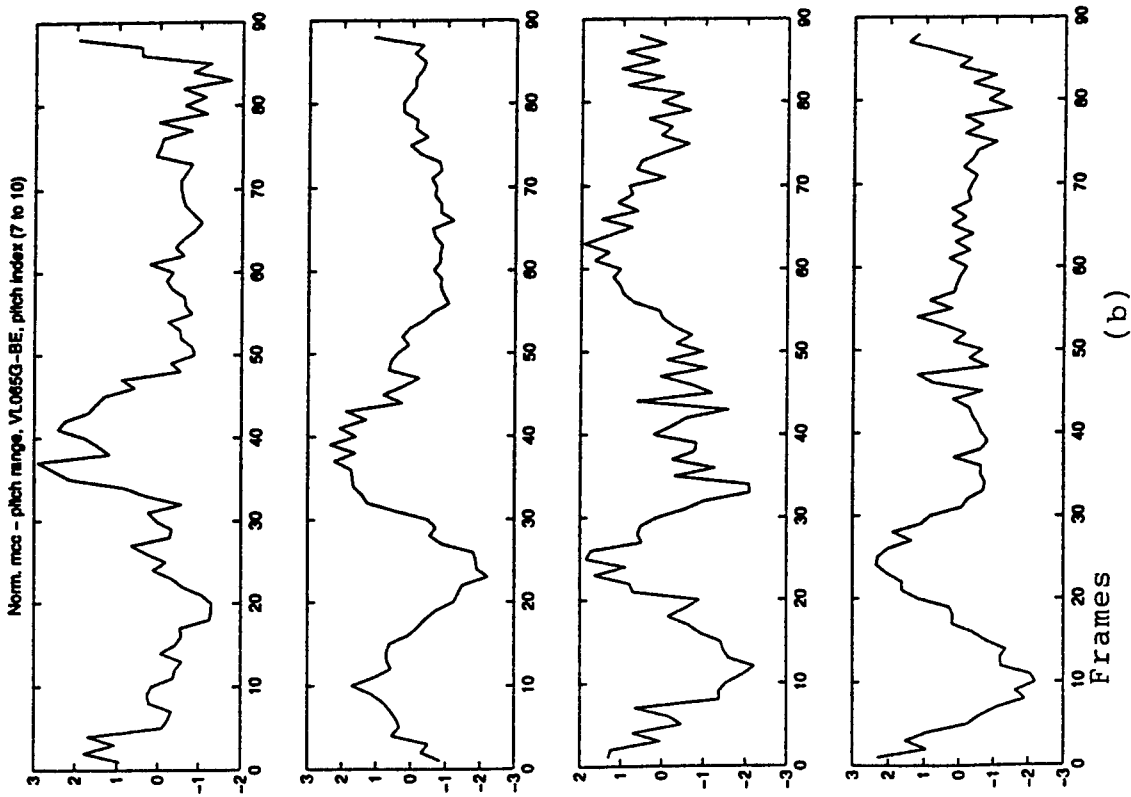
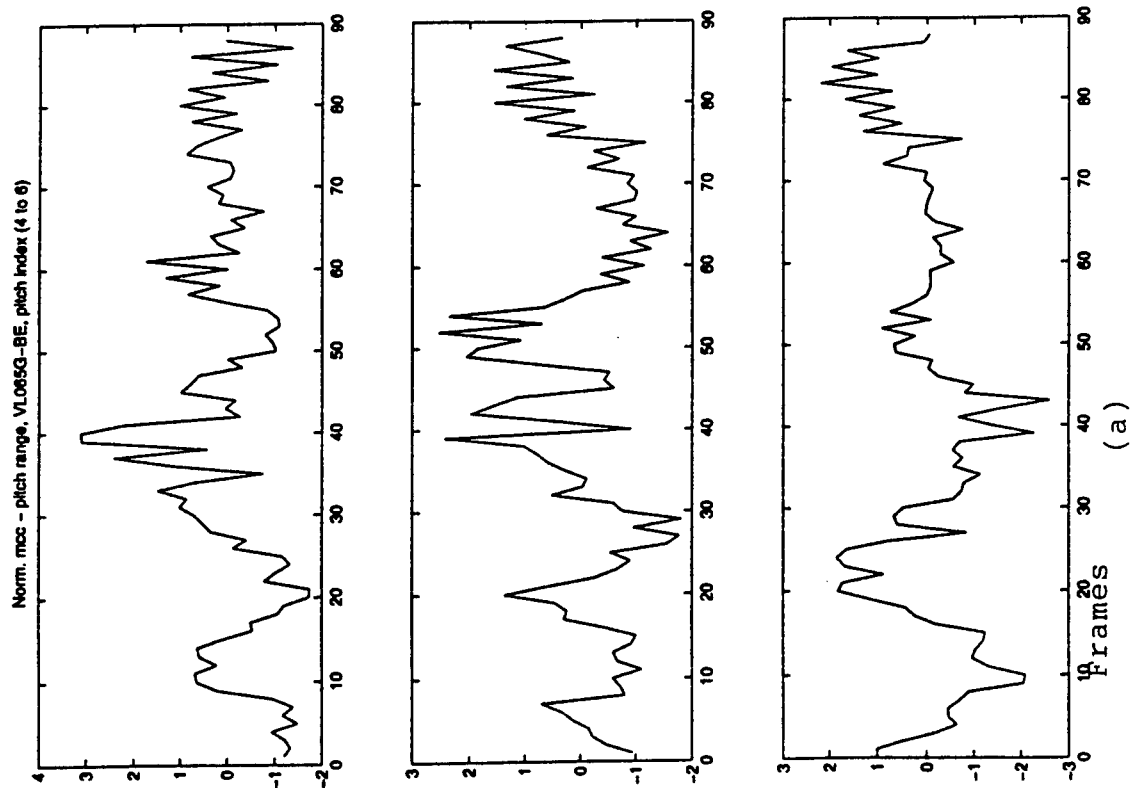


Fig.1 Mel-Cepstra in the pitch (100 Hz - 160 Hz) range for Speaker VL. (a) and (b) for high heart rate (VL065G)

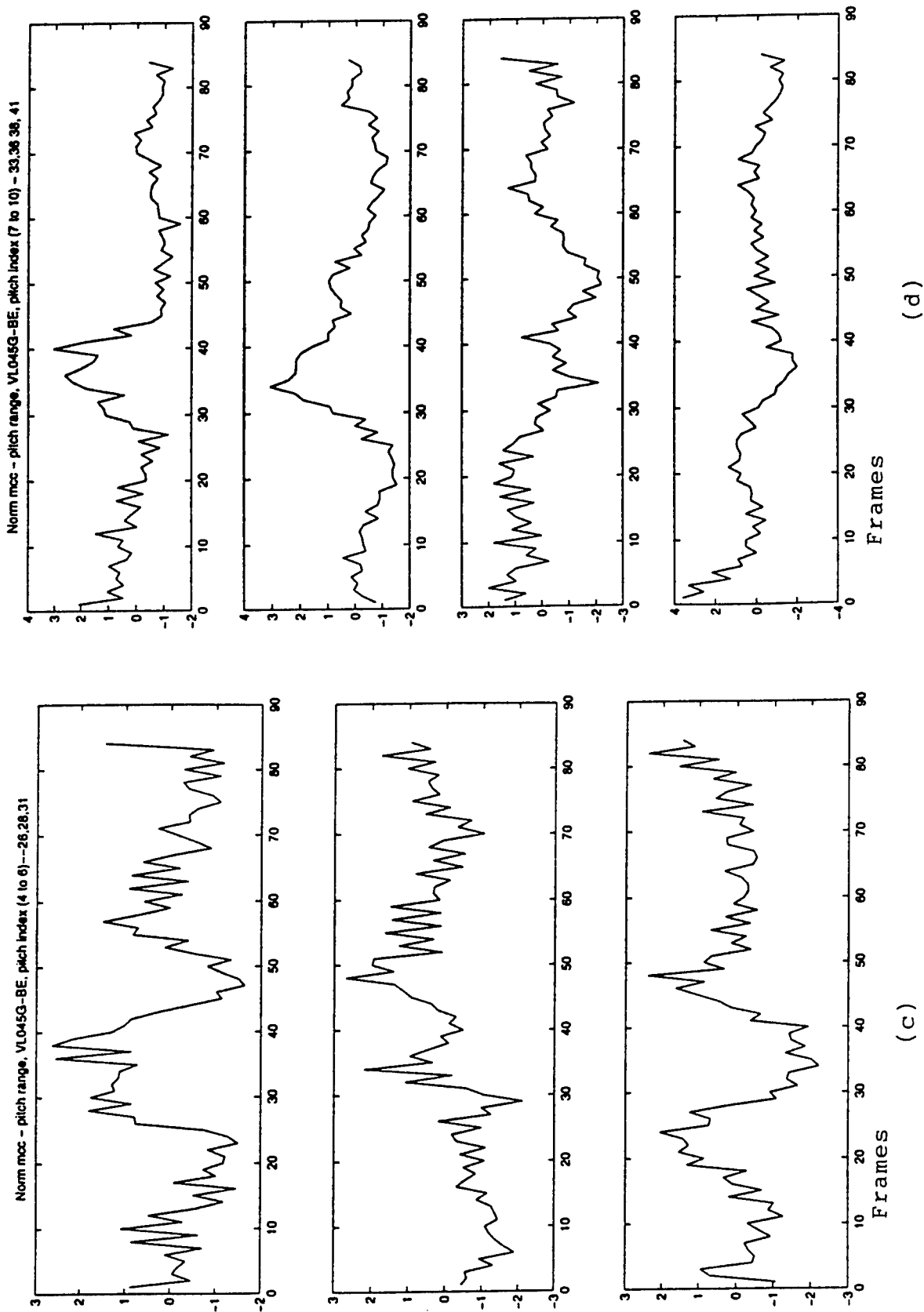


Fig.1 Mel-Cepstra in the pitch (100 Hz - 160 Hz) range for Speaker VL. (c) and (d) for medium heart rate (VL045G)

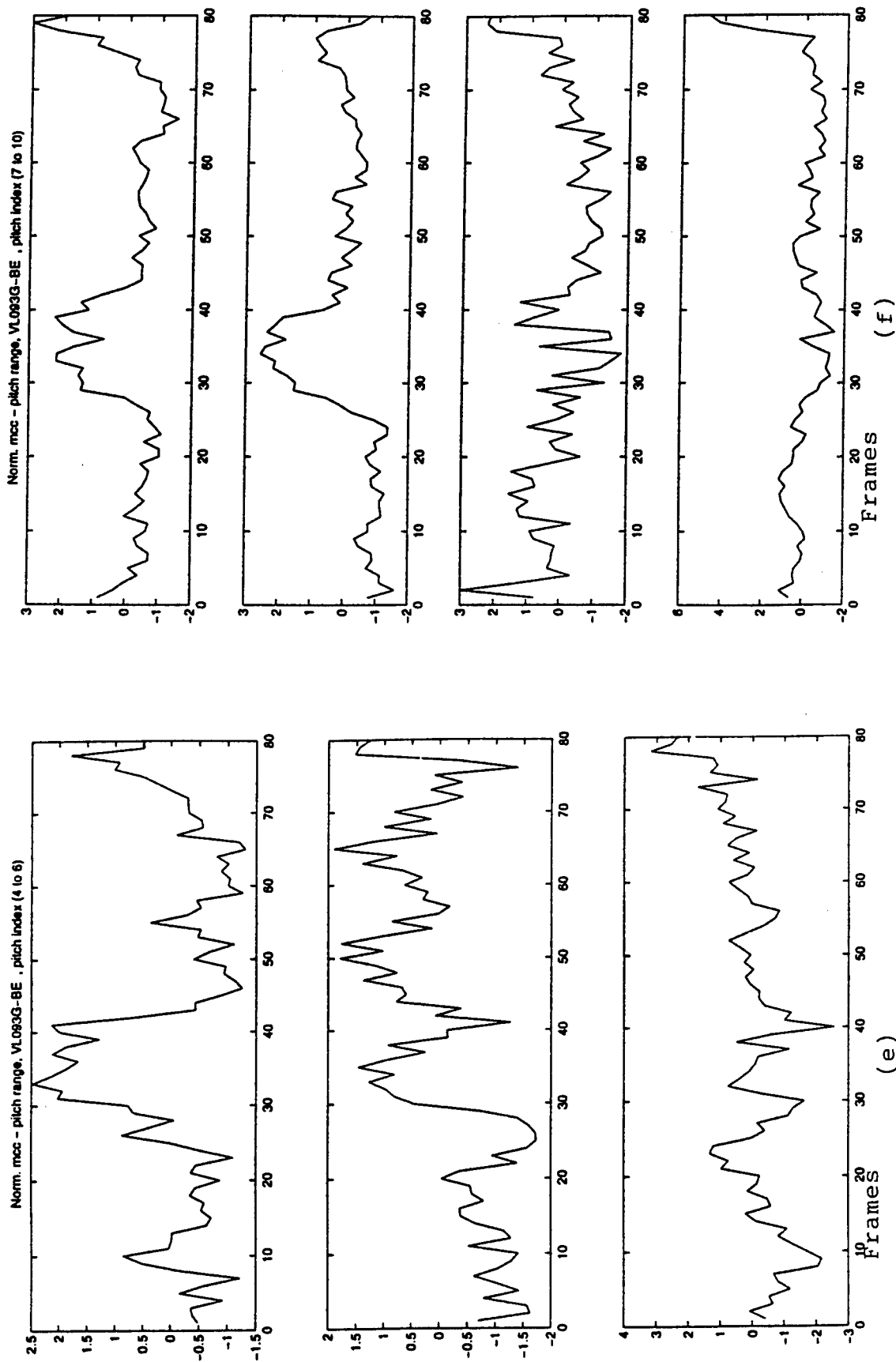


Fig.1 Mel-Cepstra in the pitch (100 Hz - 160 Hz) range for Speaker VL. (e) and (f) for low heart rate (VL093G)

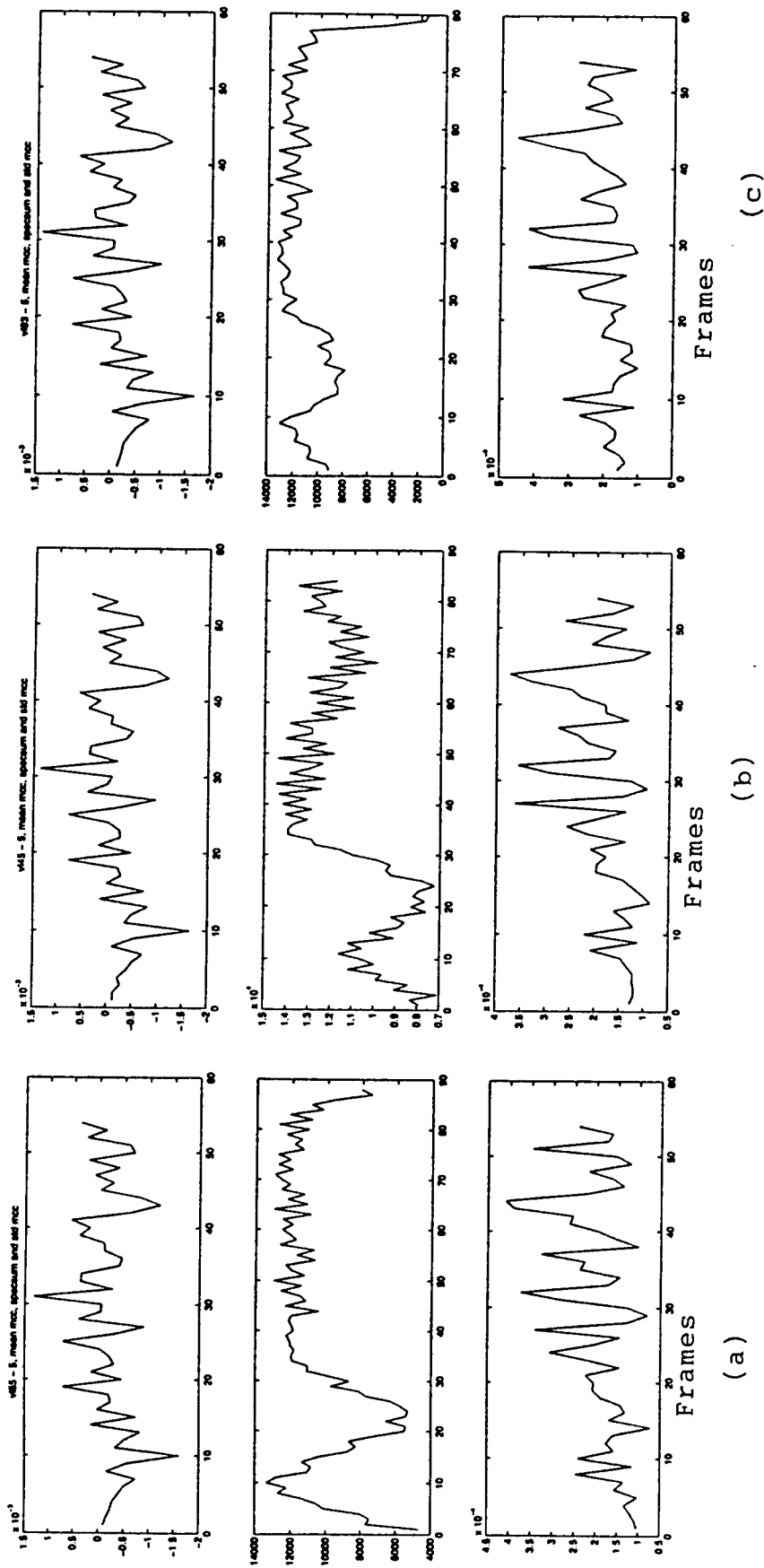


Fig. 2 Cepstral mean, total logged spectrum and standard deviation for Speaker VL for (a) high heart rate (VL065G), (b) for medium heart rate (VL045G), and (c) for low heart rate (VL093G)

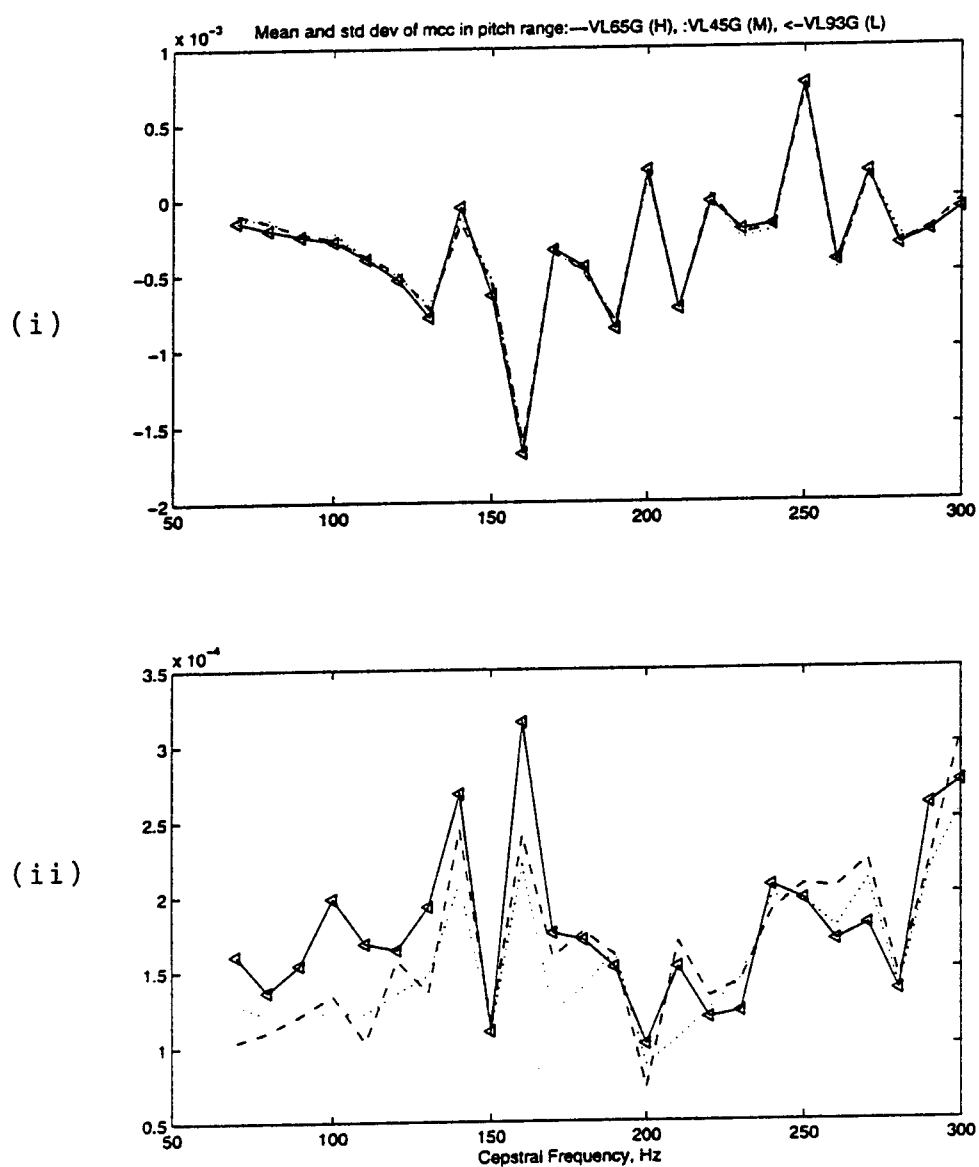


Fig. 3 (i) Mean, and (ii) standard deviation of cepstral coefficients for speaker VL at high (--), medium (..), and low (V) heart rates

Table III
Range of F0 at Measured Heart rates for speaker RD

| Index | Pitch Range, Hz | Heart Rate | Norm. Heart Rate |
|--------|--------------------|------------|---------------------|
| RD055G | 159 - 273 | 77.29 | 0.215 |
| RD021G | 101 - 181 | 76.49 | 0.108 |
| RD041G | 122 - 199 | 72.27 | -0.454 |

That the logged spectral energy varies distinctly based on the heart rate is seen in Fig. 2 (middle plot), which shows a small valley at the beginning of "eye" in "bull's eye". The valley becomes more prominent as the heart rate increases, although the peak logged spectral energy is approximately the same in all three cases. The same type of behavior in the logged spectral energy is seen for speaker RD as shown in Fig. 5. For this speaker, however,

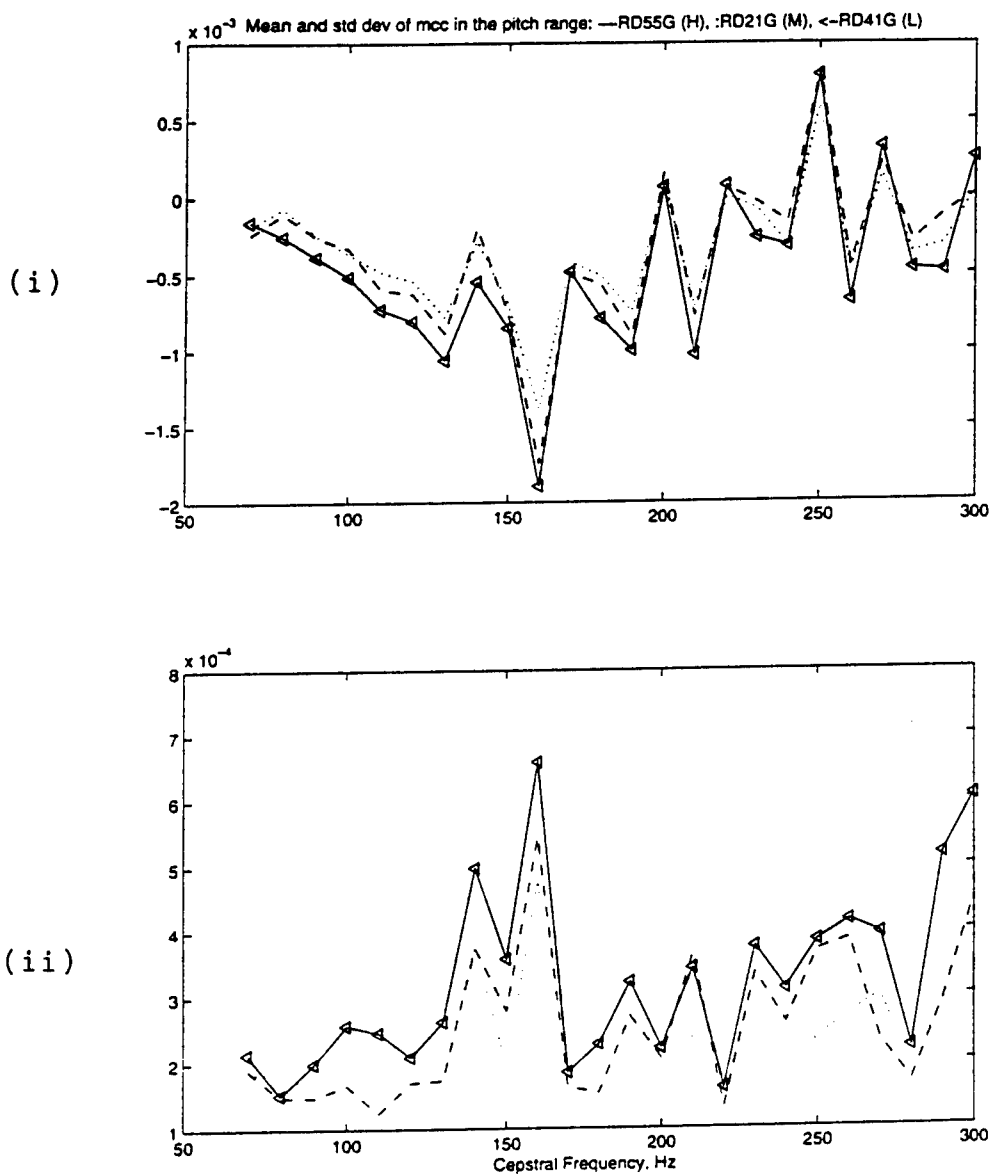


Fig. 4 (i) Mean, and (ii) standard deviation of cepstral coefficients for Speaker RD at high (--), medium (..), and low (▽) heart rates

the variation between medium (RD021G) and low (RD041G) is not as distinct as between high (RD055G) and the other two. This may be attributed to the fact that the low heart rate considered may be below the "normal" rate, as seen from the normalized rate of -0.454; hence the changes in the logged spectral energy for the lower-than-normal heart rate differs from that for the higher-than-normal heart rates. In addition, unlike for speaker VL, the peak energy appears to be in direct proportion to the heart rate. Based on the observations for these two speakers, therefore, the variations in the total spectral energy with heart rate may be hypothesized to occur at different frames and phonemes for different speakers.

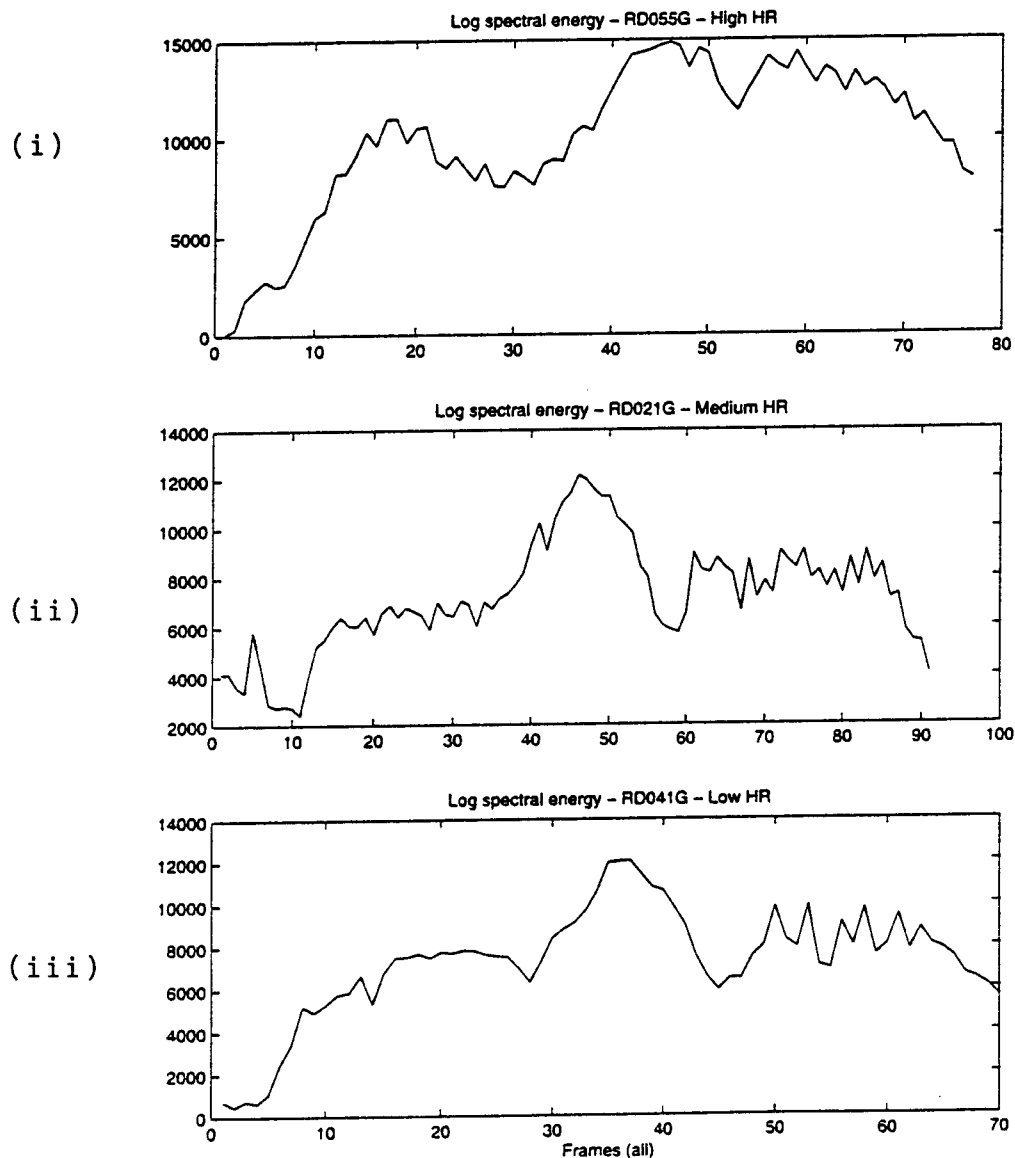


Fig. 5 Log spectral energy for speaker RD at (i) high, (ii) medium, and (iii) low heart rates

The standard deviations of the scaled cepstral coefficients at low-to-mid indices (corresponding to frequencies of 60 Hz to 180 Hz), in general, appear to be the largest at low heart rate. This has been observed for all the speakers listed in Table I. The lowest deviations at many of the cepstral indices (frequencies), however, resulted at medium heart rate instead of at high heart rate, as seen in Figs. 3 and 4. This anomaly for the smallest deviations in the cepstral values in the pitch range of indices was seen for three of the five speakers considered. As for the common behavior for all the five speakers, the largest variation in the standard deviation occurred in the range of approximately 120 Hz to 170 Hz (corresponding to cepstral index 31 to 44), regardless of the nominal pitch frequency and its variation with heart rate. This is a noteworthy feature since the pitch of the speakers in the utterances considered (Table I) varied by a minimum of 55 Hz (for VL) to a maximum of 172 Hz (for RD).

The distribution of the cepstral coefficients with time (frames) does not appear to have the same variations with heart rate for different speakers. The cepstrum at index 4 (C4, or frequency 100 Hz) for speaker VL, for example, has less fluctuations with decreased heart rate. The same type of smoothing effect with decreased heart rate is observed in Fig. 1 at other indices in the pitch range as well. Although not as smooth, similar variations occurred for speaker RD in the entire pitch range of frequencies. For other speakers, however, the fluctuations in the cepstral values increased as the heart rate decreased.

To get a better relationship between heart rate and the cepstral values, correlation of the normalized cepstral coefficients was carried out. Hidden periodicities in the coefficients were brought out in the autocorrelation of the coefficients as seen in Fig. 6 for speaker VL. At index 4 (corresponding to 100 Hz), for example, the autocorrelation of C4 at moderate heart rate shows a distinct peak at approximately frame 30 and again at frame 60. This indicates a period of approximately 30 frames, or 160 ms. At heart rates above and below, however, the peaks are either not as distinct or are at different locations for C4. Similar differences in the periodic variation of the cepstral values are observed at other indices as well. Correlation of the cepstral coefficient at index 6, which corresponds to 120 Hz, shows a period of approximately 160 ms (30 frames) at high heart rate (VL065G), and 135 ms (24 frames) at medium heart rate (VL045G); at low heart rate (VL093G), the period appears to be about 85 ms (15 frames), based on the clear peaks after frame index 40 in Fig. 6 (c).

For speaker RD, the autocorrelation of cepstral coefficients, indicates different variations with heart rate depending on the choice of the cepstral index. For index 2 (80 Hz), for example, the correlation has a period of 110 ms (20 frames) at high heart rate (RD055G), 160 ms (30 frames) at moderately high heart rate (RD021G), and 110 ms (20 frames) again at low heart rate (RD041G). This is inconsistent with the behavior observed for speaker VL. At index 5 (110 Hz), however, the periodicities decrease with decreasing heart rate: 260 ms (50 frames) at high (RD055G), 235 ms (45 frames) at medium (RD021G), and 85 ms (15 frames) at low (RD041G). As with speaker VL, these periods, in general, are consistent with decreasing heart rate at the cepstral indices considered. Although not shown, the same type of periodic variation of the autocorrelation of cepstral coefficients was observed for all the speakers studied. We note that even allowing for variable lengths in the utterances due to stress, the decrease in the autocorrelation period is consistent with decreasing heart rate.

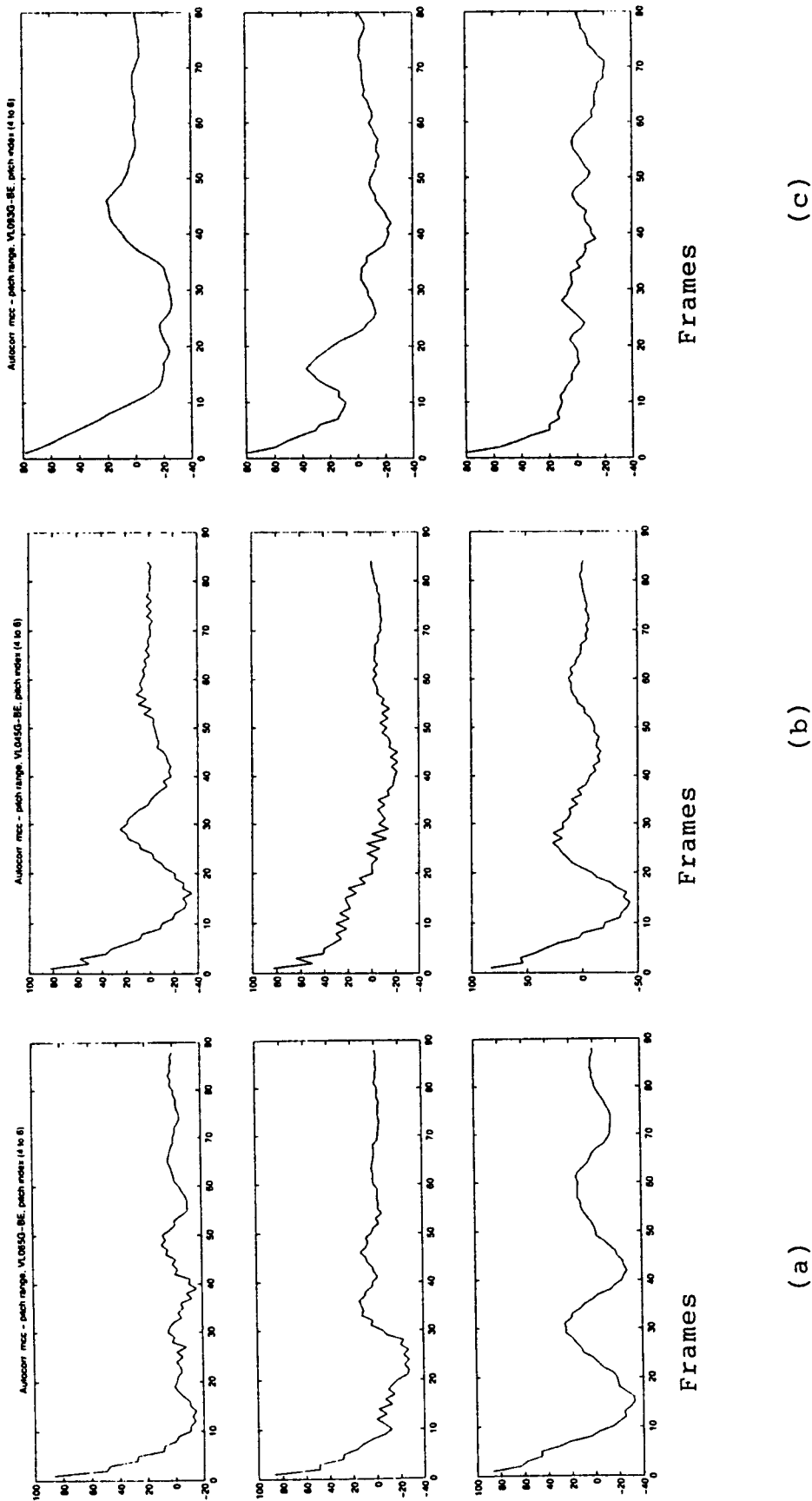


Fig. 6 Autocorrelation of cepstral values in the pitch range for speaker VL
 (a) high heart rate (VL065G), (b) medium heart rate (VL045G), and (c) low heart rate (VL093G)

For further evidence of the variation of the cepstral values with heart rate, cross correlation of the values at a few indices was performed. Fig. 7 shows the cross correlation of the cepstra at three frequencies for speaker VL. In Fig. 7a, the cross correlation between C4 (100 Hz) and C9 (150 Hz) for speaker VL shows four peaks at high heart rate, one major peak at moderate heart rate, and two major peaks at low rate. The common peak at frame index 30-35 for high and medium heart rates indicates that the cepstra at 100 Hz and 150 Hz are highly correlated around the first third of the utterance; at low heart rate, however, C4 and C9 seem to be negatively correlated. At frequencies 150 Hz (C9) and 200 Hz (C14) (Fig. 7b), the common peak between high and medium heart rates occur in the middle of the utterance while for low heart rate, no well-defined peak appears. Considering cepstra at 100 Hz (C4) and 200 Hz (C14) (Fig. 7c), a high degree of correlation appears at the beginning (around frame 10) and again in the middle of the utterance for high and medium heart rates. At low heart rate, only the middle of the utterance shows a large degree of correlation between 100 Hz and 200 Hz. Based on the three frequencies considered, it appears that, depending on the heart rate, a choice of cepstral indices may result in a better correlation of cepstra at different frequencies.

For speaker RD, cross correlation of the cepstral values at low frequencies (between 100 Hz and 150 Hz,) showed almost identical variations, indicating that the cepstra C4 and C9 vary similarly at all heart rates. Although similar identical variations were seen for 150 Hz and 200 Hz, a much stronger positive correlation appeared at mid point of the utterance at medium heart rate compared to those at high and low heart rates. Between 100 Hz and 200 Hz, the cross correlation did not indicate a very distinguishing feature among the three heart rates considered. As with speaker VL, therefore, the choice of cepstral indices appears to govern the degree of correlation among cepstra at different heart rates. The same general trend in the cross correlation of cepstral values at the three frequencies used was observed in the other speakers considered.

V. Cepstral Features in the Formant Range of Frequencies

To study the acoustical attributes of stress in the formant range of frequencies, cepstral values in the high frequency range were computed for the utterance, "*bull's eye*" spoken by the fighter controllers at various heart rates. Because of the changing formants and their bandwidths due to stress, a range of frequencies from approximately 250 Hz to 4932 Hz, with increasing bandwidths at each frequency, were chosen for cepstral computation. Table IV lists the approximate frequencies and the triangular bandwidths used.

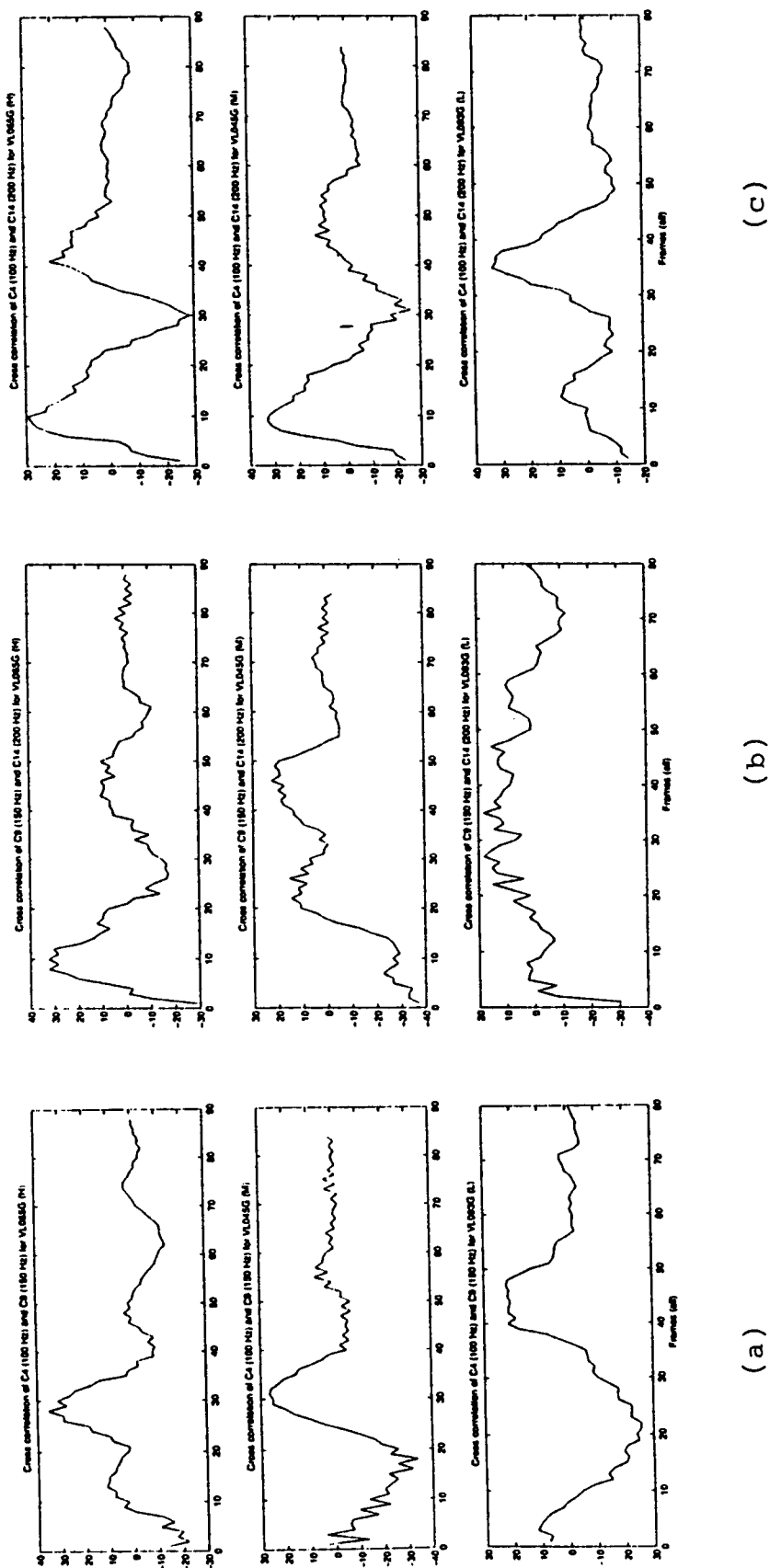


Fig. 7 Cross correlation of cepstral values at 100 Hz, 150 Hz and 200 Hz for speaker VL

(a) C4 (100 Hz) and C9 (150 Hz), (b) C9 (150 Hz) and C14 (200 Hz)

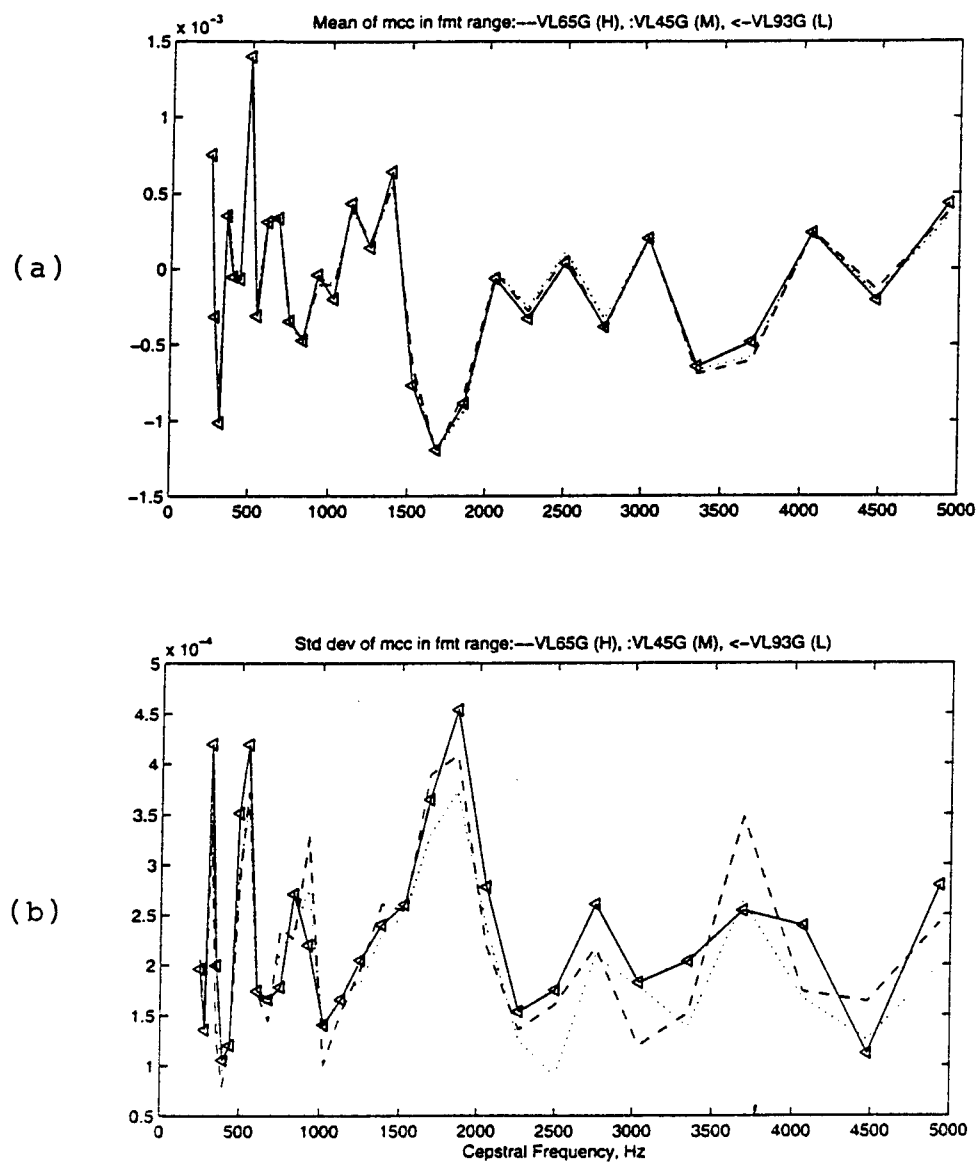
Table IV
Frequencies and Bandwidths used for High-Frequency Cepstra

| | | | | | | | | | | |
|---------------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| Index | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| Frequency, Hz | 250 | 281.5 | 316.1 | 354.3 | 396.2 | 442.3 | 493 | 548.8 | 610.2 | 677.8 |
| Bandwidth, Hz | 30 | 33 | 36.3 | 39.9 | 43.9 | 48.3 | 53.1 | 58.5 | 64.3 | 70.7 |
| Index | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
| Frequency, Hz | 752 | 833.7 | 923.6 | 1022.5 | 1131.2 | 1250.8 | 1382.4 | 1527.2 | 1686.4 | 1861.5 |
| Bandwidth, Hz | 77.8 | 85.6 | 94.2 | 103.6 | 113.9 | 125.3 | 137.8 | 151.6 | 166.8 | 183.5 |
| Index | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 |
| Frequency, Hz | 2054.2 | 2266.1 | 2499.2 | 2755.6 | 3037.7 | 3347.9 | 3689.2 | 4064.6 | 4477.6 | 4931.9 |
| Bandwidth, Hz | 201.8 | 222 | 244.2 | 268.6 | 295.5 | 325 | 357.5 | 393.3 | 432.6 | 475.9 |

As with the low quefrequency cepstra, the mean cepstral values (Figs. 8a and 9a) at the higher end of quefrequency are almost invariant with high, medium and low heart rates. The standard deviations (Fig. 8b and 9b) also show behavior similar to those at low quefrequencies: cepstra at low heart rates have, in general, higher deviations than those at higher heart rates. Unlike the cepstra in the pitch range, the deviations at low heart rate show an increase over the high and medium heart rates at almost all frequencies considered. However, the difference in the deviations between medium and high heart rates is not significant or consistent across the cepstral indices. Depending on the speaker and the heart rate difference, the difference in the deviations also vary. This type of behavior with high standard deviations at low heart rates was observed for all the speakers considered.

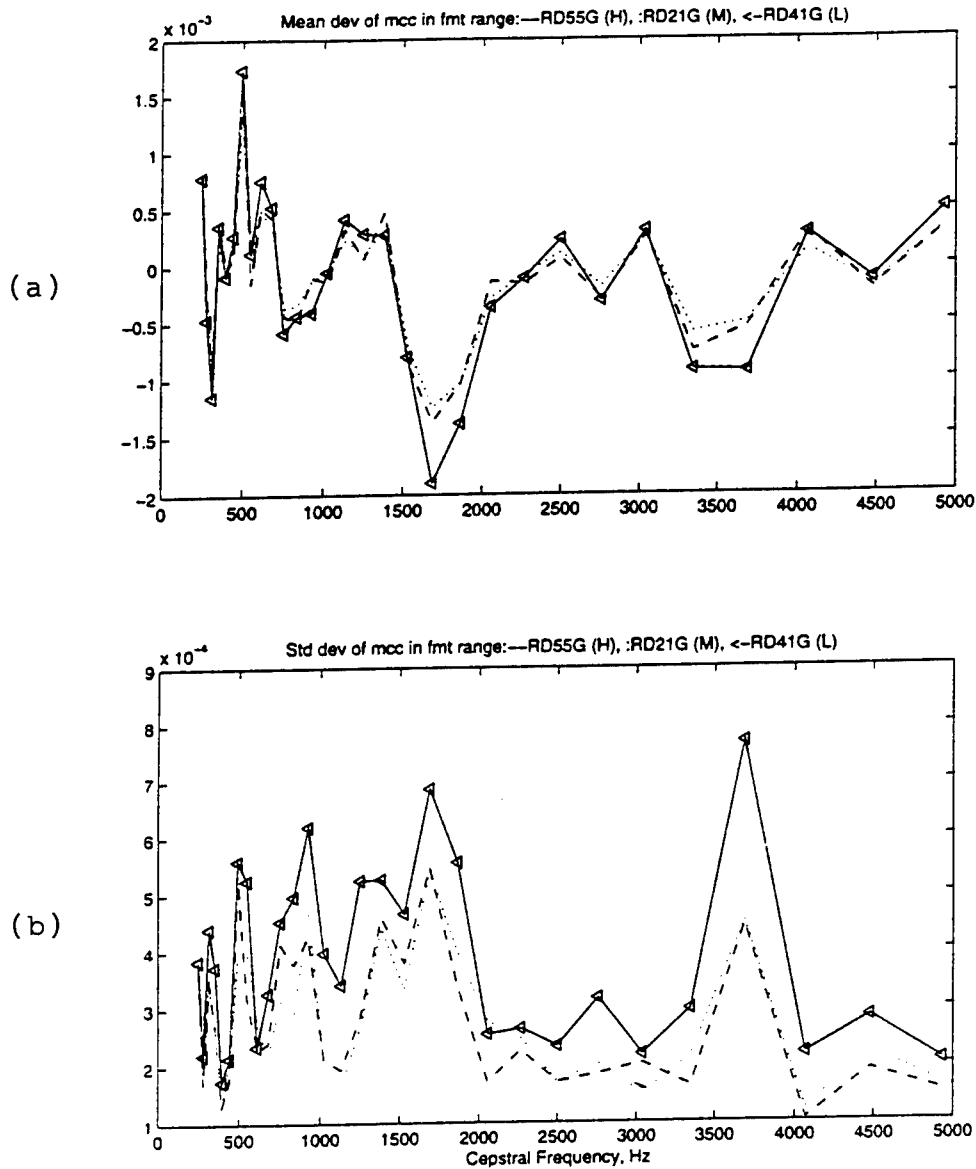
Of the 30 indices of cepstral values, four indices at 4, 14, 18 and 23 were chosen for correlation. These indices correspond to 354.3 Hz, 1022.5 Hz, 1527.2 Hz, and 2499.2 Hz with bandwidths of 39.9 Hz, 103.6 Hz, 151.6 Hz, and 244.2 Hz respectively. We note that these frequencies, except for 1022.5 Hz, are approximately in the mid range of the first three formants. The cepstral values at these frequencies do not seem to indicate any discernible pattern of behavior with heart rates.

Autocorrelation of the cepstra at the selected indices, as with that at low frequencies, shows a periodic behavior. Fig. 10 depicts the autocorrelation at the four frequencies for speaker VL. The period in the autocorrelation at 354.3 Hz shows an increase with decreasing heart rate: 28 frames at high, 30 frames at medium and 35 frames at low. Similar increase in the period can be seen at 2499.2 Hz. This increase is in contrast with the decreasing period in the autocorrelation of cepstra at low frequencies (Fig. 6). At the two other chosen frequencies, a well-defined periodic variation is not observed.



Figs. 8(a) Mean and (b) standard deviation of cepstral values at the high end of frequencies at high (--), medium (...), and low (∇) heart rates for speaker VL

The above behavior was also observed in the autocorrelation of the cepstra for speaker RD. For this speaker, however, the low heart rate does not yield a proportionally longer periods at the chosen cepstral indices. This again demonstrates that the stress-related changes in the cepstra occur at different indices for different speakers.



Figs. 9 (a) Mean and (b) standard deviation of cepstral values at the high end of frequencies at high (--), medium (...), and low (▽) heart rates for speaker RD

Cross correlation of the cepstra at the selected indices (Fig. 11) shows a smoothing trend with fewer peaks as the heart rate decreases from high to medium to low. This smoothing behavior was consistently seen at other frequencies as well. In addition, the same trend was observed for each of the speakers considered.

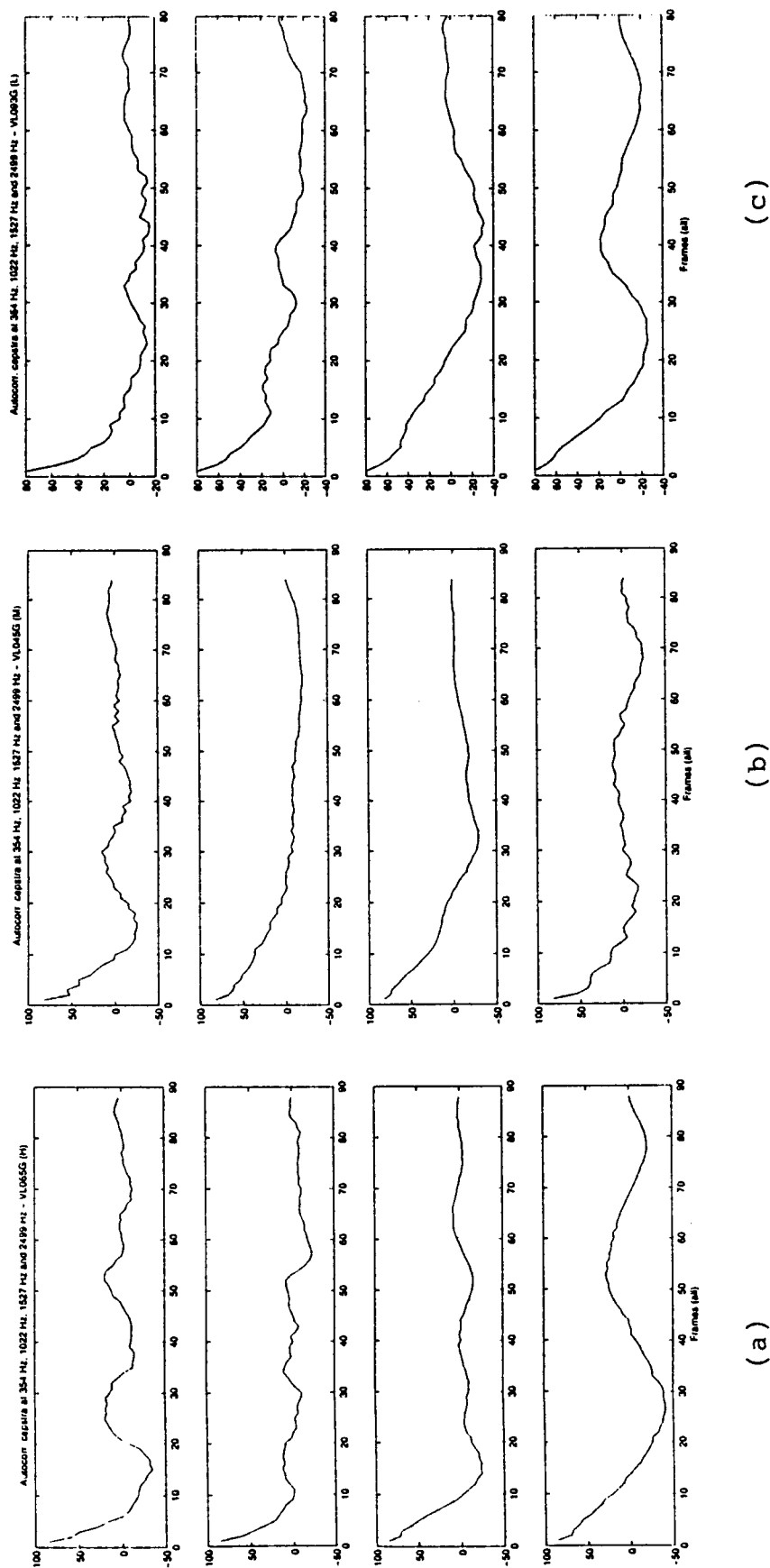


Fig. 10 Autocorrelation of cepstral values at (i) 354.3 Hz, (ii) 1022.5 Hz, (iii) 1527.2 Hz, and (iv) 2499.2 Hz for speaker VL (a) high heart rate (VL065G), (b) medium heart rate (VL045G), and (c) low heart rate (VL093G)

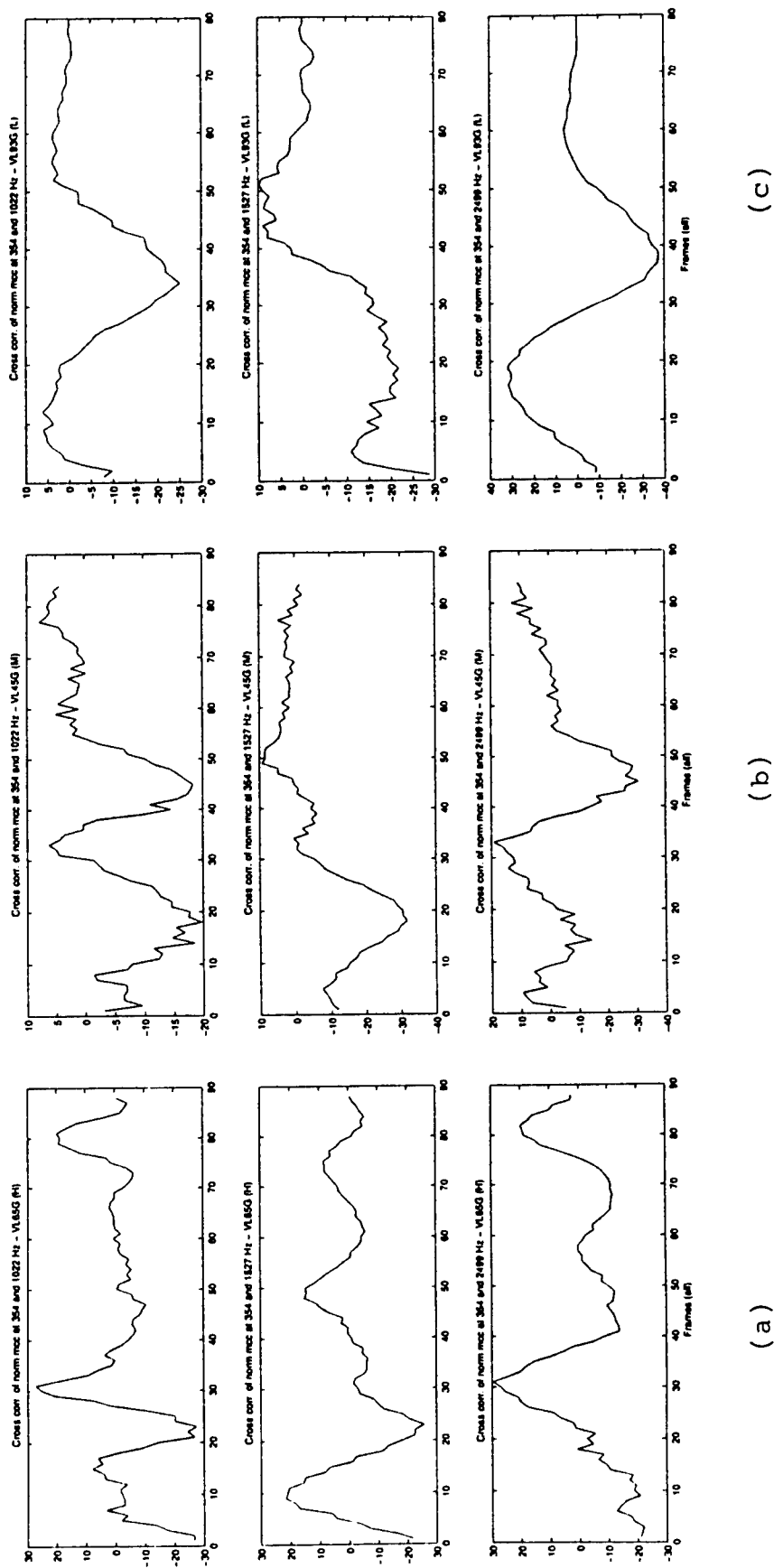


Fig. 11 Cross correlation of cepstral values at selected indices: (i) 354.3 Hz and 1022.5 Hz, (ii) 354.3 Hz and 1527.2 Hz, and (iii) 354.3 Hz and 2499.2 Hz for speaker VL

(a) high heart rate (VL065G), (b) medium heart rate (VL045G), and (c) low heart rate (VL093G)

VI. Discussion

The present work analyzed features based on the gross variations of mel-distributed cepstra at selected indices to determine any correlation with a speaker's heart rate. From the different parameters considered using low- and high-frequency cepstra, it appears that the cepstral values themselves do not seem to indicate a direct relationship to the speaker's heart rate, at least not at the indices chosen. The mean cepstral value at each index across an utterance is virtually constant for a speaker regardless of his heart rate. While this feature is not useful in analyzing stress based on heart rate, it can be applied as a robust feature in speaker and speech recognition applications.

The standard deviation of the cepstral values, on the other hand, varies inversely with heart rate. Regardless of the pitch variation for a speaker, the largest variation of the standard deviation was observed to occur in a narrow range of quefrency. Similar to pitch, therefore, the standard deviation of cepstral values in the range of 120 Hz to 170 Hz can be used to correlate with heart rate. At the higher quefrency scale (corresponding to above 1500 Hz), in general, the standard deviations appear to be much larger than those at lower quefrequencies (from 250 Hz to 1500 Hz). Based on the two sets of deviations, therefore, cepstral deviations in a narrow (pitch) range of frequencies are relatively more efficient than those in the formant range in comparing relative heart rates. Pitch-synchronous analysis of cepstra at selected phonemes is more likely to reflect the changes in the pitch due to heart rate changes. More data at higher indices and lower bandwidths may be needed to conclusively verify the comparative levels of heart rate using formant range of frequencies.

Although results of using Hamming window were reported here, other windows, in particular, exponential [13, 14] and triangular windows were also used in the early part of this research. Exponential window was applied to the cepstral analysis to mimic the auditory processing [13] so that stress detection may be reflected in the observed cepstra. Results using a 15 ms (240 sample) exponentially rising window with a time-constant of 4 ms did not show any improvement over those using Hamming window. Figs. 12 and 13, for example, show the mean and the standard deviations of cepstra in the pitch and formant range for speaker VL.

Peak log spectral energy is seen increasing with heart rate. The location of the peak, however, is different for different speakers. Based on a number of speakers and different phonemes, the peak energy feature may be calibrated for proportionality with heart rate.

Correlation of the cepstra at the selected indices, however, show the following trends with high, medium and low heart rates. At low cepstral indices, autocorrelation of cepstra for a speaker show peaks occurring at different intervals dependent on the relative heart rates. Autocorrelation of cepstra at low quefrequencies have peak intervals decreasing with decreased heart rates. The extent of decrease in interval, however, depends on the choice

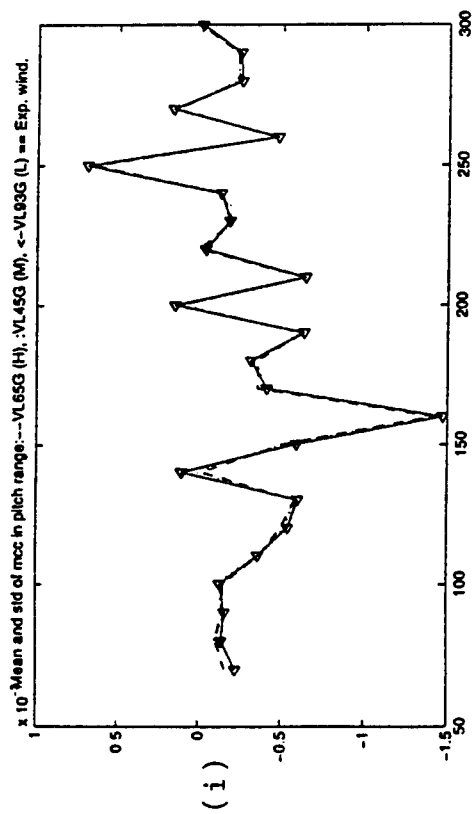
of the cepstral index for a given speaker. This is akin to the dependence of energy in a spectral band on the stress: if the band is centered at nominal pitch frequency, the energy variation corresponds to the shift of pitch due to stress. As with cepstra, therefore, a speaker's nominal pitch range may be helpful in narrowing the range for spectral shift determination. At the selected higher end of indices, the peak intervals show an increase with decreasing heart rate. Here again, the behavior may be different with different indices. Also, based on the bandwidths used at each index and depending on the location and bandwidth of formants, the change in cepstra with heart rate, and hence, the autocorrelation could be different from that observed here.

Cross correlation of cepstra at selected frequencies, such as the formants at normal heart rates for a speaker, is more likely to yield large variations at elevated heart rates. These variations may also reflect in the cepstral values and their standard deviations at the selected indices. The choice of center frequencies and bandwidths for the cepstra must accommodate possible shifts in both due to stress. With data for the resting heart rate of a speaker and corresponding utterance, correlation and standard deviation of cepstra can yield values proportional to the spectral shifts.

VII. Conclusion

In this project, variation of a set of cepstral parameters for speech spoken at different heart rates was studied. For comparison, the utterance for "*bull's eye*" spoken by male European flight controllers was considered in each case. It was found that the choice of cepstral indices determined the degree of variation of the cepstra and their correlated values with a speaker's heart rate. As with pitch, these indices varied for each speaker. If the cepstral indices are in the vicinity of nominal pitch frequencies, the cepstra, their correlation, and the standard deviations all showed proportional variation with heart rate. Cepstra at higher frequencies also resulted in variations proportional to heart rate when the indices and their bandwidths corresponded to nominal (unstressed) values. Mean value across an utterance for each speaker at each cepstral index showed an almost constant value independent of heart rate. Hence, the mean cepstral value at one or more indices could be used as a robust feature in speech and speaker recognition applications.

Further work using pitch-synchronous cepstral analysis at speaker-dependent indices is expected to yield a more measurable correlation of cepstral features with heart rate. Once this correlation is established, deviation of the features for different phonemes may be studied at different heart rates. This study can further lead to the analysis of a speaker's stress-induced heart rate variation using speech-independent cepstral features.



10-25

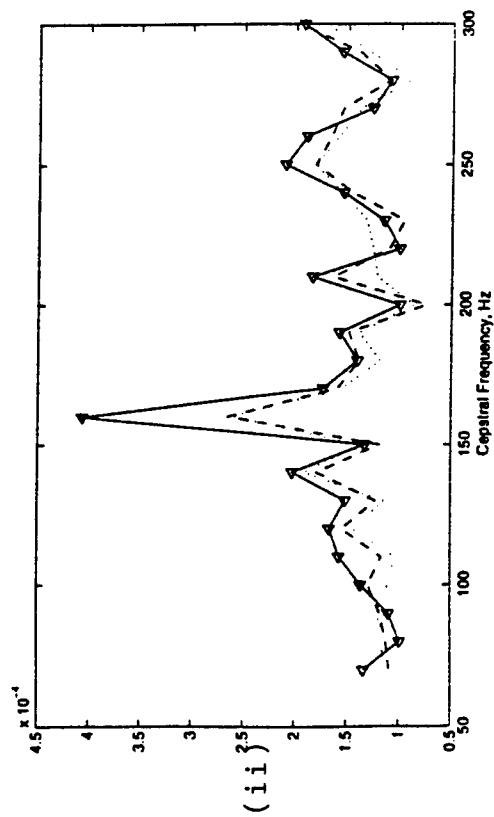


Fig. 12 (i) Mean, and (ii) standard deviation of cepstral coefficients in pitch range using exponential window for speaker VL at high (---), medium (.), and low (V) heart rates

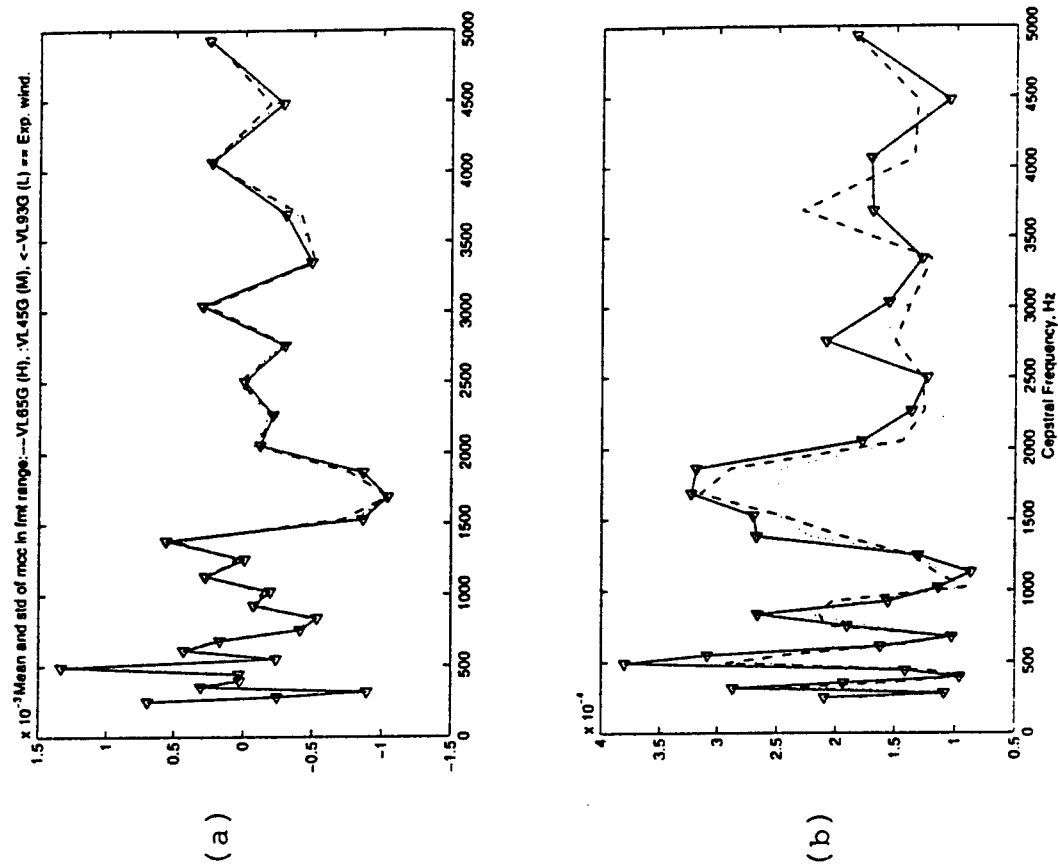


Fig. 13 (a) Mean and (b) standard deviation of cepstral values at the high end of frequencies using exponential window at high (---), medium (.), and low (V) heart rates for speaker VL

References

- [1] Lieberman, P. and S.B. Michaels, "Some Aspects of Fundamental Frequency and Envelope Amplitude as related to the Emotional Content of Speech," *Acoust. Soc. Am.* Vol. 34, No. 7, 1962, pp. 922-927.
- [2] Williams, C.E. and K.N. Stevens, "Emotions and Speech: Some Acoustical Correlates," *J. Acoust. Soc. Am.* Vol. 52, No. 4 (Part 2), 1970, pp. 1238-1250.
- [3] Levin, H. and W. Lord, "Speech Pitch Frequency as an Emotional State Indicator," *IEEE Trans. Systems, Man, Cybernetics*, Vol. SMC-5, No.2, March 1975, pp. 259-273.
- [4] Streeter, L.A., et al, "Acoustic and Perceptual Indicators of Emotional Stress," *J. Acoust. Soc. Am.* Vol. 73, No. 4, 1983, pp. 1354-1360.
- [5] Cairns, D.A. and J.H.L. Hansen, "Nonlinear Analysis and Classification of Speech under Stressed Conditions," *J. Acoust. Soc. Am.* Vol. 96, No. 6, 1994, pp. 3392-3400.
- [6] Protopapas, A. and P.Lieberman, "Effects of Vocal F0 Manipulations on Perceived Emotional Stress," *Proc. NATO Workshop on Speech Under Stress*, 1996, pp. 1-4.
- [7] Hansen, .H.L and B.D. Womack, "Feature Analysis and Neural Network-based Classification of Speech under Stress," *IEEE Trans. Speech Audio Processing*, Vol. 4, No. 4, pp. 307-313, July 1996.
- [8] Rajasekaran, P.K., G.R. Doddington and J.W. Picone, "Recognition of Speech under Stress and in Noise," *Proc. ICASSP86*, 1986, pp. 733-736.
- [9] Chen, Y., "Cepstral domain Stress Compensation for Robust Speech Recognition," *Proc. ICASSP87*, 1987, pp. 717-720.
- [10] Stanton, B.J., L.H. Jamieson and G.D. Allen, "Acoustic-Phonetic Analysis of Loud and Lombard Speech in Simulated Cockpit Conditions," *Proc. ICASSP88*, 1988, pp. 331-334.
- [11] Stanton, B.J., L.H. Jamieson and G.D. Allen, "Robust Recognition of Loud and Lombard Speech in the Fighter Cockpit Environment," *Proc. ICASSP89*, 1989, pp. 675-678.
- [12] Bou-Ghazale, S.E. and J.H.L. Hansen, "Duration and Spectral based Stress Token Generation for HMM Speech Recognition under Stress," *Pr. ICASSP94*, 1994, Vol. II, pp. 45-49.
- [13] Rhody, H.E., R.A. Houde, C.W. Parkins and S. Dianat, "Speech Analysis based on a Model of the Auditory System," *Rome Laboratory Report*, 1985, Rome, NY.
- [14] B.L. Losiewicz, "Windowing Comparison Project: The Effect of Window Shape and Size on Phoneme Identifiability," *Rome Laboratory Report*, 1993, Rome, NY.
- [15] Waves+ Manual, Entropic Research laboratory, Inc., Washingto, DC, 1996.

MODEL ORDER SELECTION
FOR MULTICHANNEL
INNOVATIONS BASED DETECTION
IN AIRBORNE RADAR

by

Julio Castro and James P. LeBlanc

Klipsch School of ECE
New Mexico State University
Dept 3-0, Box 30001
Las Cruces, NM 88003

Final Report for:
Summer Research Extension Program

June 3, 1998

SUPPORTED BY THE AIR FORCE OFFICE OF SCIENTIFIC RESEARCH, BOLLING AFB AND NEW MEXICO STATE UNIVERSITY

MODEL ORDER SELECTION FOR MULTICHANNEL INNOVATIONS BASED DETECTION IN AIRBORNE
RADAR *

Julio Castro

julcastr@nmsu.edu

James P. LeBlanc

leblanc@nmsu.edu

Klipsch School of Elec. and Comp. Eng.

New Mexico State University

Las Cruces, NM 88001

Abstract

This paper investigates the model order selection problem for use with the multichannel autoregressive (MAR) process in airborne radar detection processing which uses an Innovations Based Detection Architecture (IBDA). Results indicate that a low order model should be used. Specifically, this paper investigates the use of the Akaike Information Criterion (*AIC*) and prediction error power over independent realizations for model order selection. Examples are included for physically modeled data sets as well as actual radar data sets.

*Supported by the United States Air Force, contract number: F49620-93-C-0063

1 Introduction

It has been established that parametric modeling for target detection is an alternative method to correlation-based detection methods [1]. Parametric modeling using autoregressive processes has been used extensively in the literature for various applications in the univariate case as well as in the multivariate case [2] [3] [4] [5]. However, literature discussing the application of multichannel autoregressive processes is relatively limited for the airborne radar target detection application [6]. Noticeably absent in the literature is guidance and recommendations for model-order selection for this application. This research investigates the model order selection and applicability of Multichannel Autoregressive (MAR) processes to the airborne radar surveillance target detection.

The use of the MAR structure to model the processes associated with the airborne radar surveillance scenario lies in the fact that this type of signals can be interpreted as a complex vector time-series signals and that the clutter can be represented as the output of a vector AR system. Initial results of such observations, presented herein as well as in [7] and [8], and show that the order demanded for such an application is relatively low. The work herein is in agreement with results showing the success of using relatively low-order models for target detection [9]. This is a critical point for airborne surveillance radar applications since detection algorithms must be feasible to implement in real-time and with limited computational resources.

In general, the J -channel multichannel autoregressive modeling consists in estimating a set of MAR parameters (complex matrices) $A(k) \in \mathbb{C}^{J \times J}, 1 \leq k \leq p, k, p \in \mathbb{N}$, and the driving noise covariance matrix $\Sigma \in \mathbb{C}^{J \times J}$, which in some sense, gives the minimum prediction error and yet satisfy the principle of *parsimony*. A MAR process can be realized as in Figure 1, where $U(n)$ is the driving noise vector with correlation matrix $R_{UU}(k)$. The complex matrices $A(k)$ represent the feedback coefficients for the vector MAR process $X(n)$.

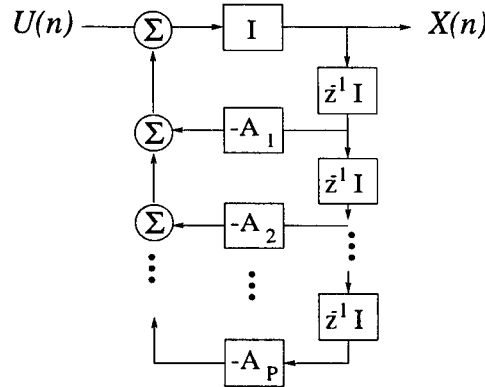


Figure 1: MAR model

An important aspect of using MAR for radar detection is the determination of the model order (p). For the most part, the selection of the model order dictates how well the model will follow the true spectrum. A model order that is too high will cause spurious peaks in the frequency response as well as a high parameter variance. A too low model order will cause unwanted smoothing of the exhibiting spectrum

[2]. Both cases (too high or too low) will adversely affect detection performance. An appropriate model order selection criterion has been a challenging formulation for MAR processes. However, several researchers have introduced reasonable criteria for model order use for specific scenarios. One of these criteria is the (*AIC*) Akaike Information Criteria, also known as An Information Criteria [10]. This research explores the use of the *AIC* and the prediction error power (PEP) for MAR model order selection in the airborne surveillance radar scenario.

Section 2 discusses in more detail the airborne surveillance radar detection problem. Section 3 presents the MAR theory. Section 4 explains the *AIC* and PEP for the multivariate case. Section 5 presents numerical results using synthetic and actual radar data.

2 Radar Target Detection

In general, the airborne surveillance radar target detection consists in determining whether a target is present or not in a large number of range cells. Additionally, if a target is present, estimating its speed and range is desired. This problem has been studied extensively since the conception of radar technology; however, new methods continue to appear in the literature. A driving force behind continued research in this problem is that there is always the need to detect the target faster, cheaper (economically and/or computationally) and more accurately as well as robustly.

A received radar return at the J element array for range cell m can be represented in matrix form as,

$$\mathbf{X}(m) = \begin{bmatrix} x_1(1) & x_1(2) & \cdots & x_1(N) \\ x_2(1) & x_2(2) & \cdots & x_2(N) \\ \vdots & \vdots & \ddots & \vdots \\ x_J(1) & x_J(2) & \cdots & x_J(N) \end{bmatrix} \quad (1)$$

where N is the number of pulses in a coherent processing interval (CPI). In general, each signal component, $x_i(n)$, at the i th array element and at time n is equal to

$$x_i(n) = s(n) + j(n) + c(n) + w(n) \quad (2)$$

where $s(n)$ is the target return reflection, $j(n)$ is a possible jammer, $c(n)$ is clutter reflection, and $w(n)$ is a white noise component.

The correlation matrix $R_{xx}(k)$ of signal x is the $J \times J$ matrix

$$R_{xx}(k) = \begin{bmatrix} r_{11}(k) & r_{12}(k) & \cdots & r_{1J}(k) \\ r_{21}(k) & r_{22}(k) & \cdots & r_{2J}(k) \\ \vdots & \vdots & \ddots & \vdots \\ r_{J1}(k) & r_{J2}(k) & \cdots & r_{JJ}(k) \end{bmatrix} \quad (3)$$

where each element $r_{ij}(k) = E\{x_i(n)^* x_j(n+k)\}$ is the correlation between array element i and j at lag k .

In addition to the classical correlation-based detection [11], approaches that try to solve this problem are the Innovation Based Detection Architectures (IBDA) , Adaptive schemes, and Parametric Adaptive Matched Filter (PAMF). The Innovation Based Detection Architectures were first presented in [12]. The experimental analysis of IBDA for surveillance radar was presented in [13]. Later and more importantly in relation to the MAR case is the Model-Based Detection method introduced by Zhang and Haykin [1]. Wherein they develop a method of target detection using the power spectrum relation that exists between the autoregressive coefficients and its frequency transform as well as exploring its use in target detection in a multichannel scenario. The difference between the MAR method and their work lies mainly in that their method deals with a stationary radar antenna as oppose to an airborne radar. Also they do not consider a whitening processing stage.

An example of an adaptive scheme is the one introduced by Brennan and Reed [14] which applied the adaptive theory popularized by Widrow [15]. However, as in any adaptive scheme, this method must take into account the convergence rate, step size considerations, processing time vs. acquisition time among many other critical conditions. For this reason modifications to this seminal paper appeared later in the literature.

In [16] a method is developed using the State Space approach. In this method, modeling and whitening is used to model ground clutter to consequently aid in target detection. Notice that all MAR models can be represented as state space models from a relation of the state and the autoregressive coefficients, however the opposite is not necessarily true [17], [18].

3 Multichannel Autoregressive Process

In general, a multichannel autoregressive process can be described by the following equation

$$X(n) = - \sum_{k=1}^p A(k)X(n-k) + U(n), \quad (4)$$

where $A(k) \in \mathbb{C}^{J \times J}$, are the complex feedback parameter matrices $X(k) \in \mathbb{C}^J$ is the data column vector, and $U(n)$ is a multichannel white input vector process whose autocorrelation function satisfies $R_{UU}(k) = \Sigma\delta(k)$, where Σ is the variance of the driving process. Notice that this type of modeling can also be viewed as an innovation process as described in [19].

In innovation-based detection architectures (IBDA) , the received signal is converted to an innovations process as shown in Figure 2.

The filtering of $X(n)$ produces an output $\varepsilon(n)$, which has been whitened temporally as,

$$\varepsilon(n) = X(n) + \sum_{k=1}^p A(k)X(n-k) \quad (5)$$

notice that $\varepsilon(n)$ may be viewed as the prediction error of the classical linear prediction problem.

The idea behind IBDA is that the prediction filter coefficients $A(k)$ would contain information about the received signal (clutter, jamming, signal), and hence a detection test could be based upon the achieved

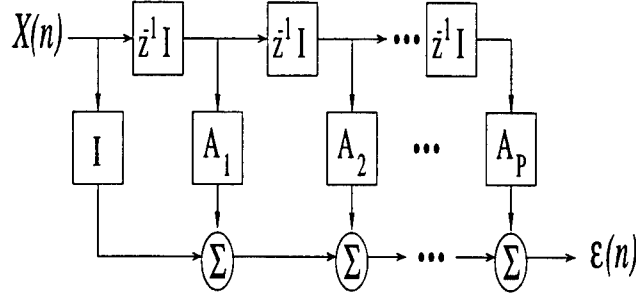


Figure 2: Innovations Process

parameterization of the $A(k)$ s.

A solution for finding the set of MAR parameters $\{A(k), 1 \leq k \leq p\}$ is by solving the multichannel Yule-Walker equations

$$\begin{bmatrix} R(0) & R(-1) & \cdots & R(-p+1) \\ R(1) & R(0) & \cdots & R(-p+2) \\ \vdots & \vdots & \ddots & \vdots \\ R(p-1) & R(p-2) & \cdots & R(0) \end{bmatrix} \begin{bmatrix} A^T(1) \\ A^T(2) \\ \vdots \\ A^T(p) \end{bmatrix} = - \begin{bmatrix} R(1) \\ R(2) \\ \vdots \\ R(p) \end{bmatrix} \quad (6)$$

where $R(k) = E\{X^*(n)X^T(n+k)\}$ and it can be shown that $R(-k) = R^H(k)$.

Then the parameters are found from correlations as follows

$$\begin{bmatrix} A^T(1) \\ A^T(2) \\ \vdots \\ A^T(p) \end{bmatrix} = - \begin{bmatrix} R(0) & R(-1) & \cdots & R(-p+1) \\ R(1) & R(0) & \cdots & R(-p+2) \\ \vdots & \vdots & \ddots & \vdots \\ R(p-1) & R(p-2) & \cdots & R(0) \end{bmatrix}^{-1} \begin{bmatrix} R(1) \\ R(2) \\ \vdots \\ R(p) \end{bmatrix} \quad (7)$$

However, lacking a closed form expression for the needed correlation matrices $R(k)$, estimated values $\hat{R}(k)$ are computed from the received signal. To ensure positive definiteness of $\hat{R}(k)$, typically the biased estimates are used. The driving covariance matrix is found as follows

$$\begin{aligned} \Sigma &= E\{X^*(n)[X(n) - \hat{X}(n)]^T\} \\ &= R(0) + E\left\{X^*(n) \sum_{k=1}^p X^T(n-k) A^T(k)\right\} \\ &= R(0) + \sum_{k=1}^p R(-k) A^T(k) \\ &= R(0) + \sum_{k=1}^p R^H(k) A^T(k) \end{aligned} \quad (8)$$

where $\hat{X}(n)$ is the forward prediction define as $\hat{X}(n) = -\sum_{k=1}^p A(k)X(n-k)$. Notice that the prediction

error power, here called *PEP*, is equal to the trace of Σ and the off diagonals elements correspond to the (spatial) cross-correlation of the driving noise.

The relation between the spectrum of the multichannel signal $X(n)$ and the two dimensional spectrum is discussed in [20], a brief formulation of this result follows.

The transfer response of a multichannel impulse response matrix $H(k)$ is given by

$$H(z) = \begin{bmatrix} H_{11}(z) & H_{12}(z) & \cdots & H_{1M}(z) \\ H_{21}(z) & H_{22}(z) & \cdots & H_{2M}(z) \\ \vdots & \vdots & \ddots & \vdots \\ H_{M1}(z) & H_{M2}(z) & \cdots & H_{MM}(z) \end{bmatrix} \quad (9)$$

where $H_{ij}(z) = \sum_{k=0}^{\infty} H_{ij}(k)z^{-k}$. For a finite time, say $k = 0, 1, \dots, N$, the matrix $H(z)$ can also be written as

$$\begin{bmatrix}
H_{11}(0) + H_{11}(1)z^{-1} \dots + H_{11}(N)z^{-N} & H_{12}(0) + H_{12}(1)z^{-1} \dots + H_{12}(N)z^{-N} & \dots & H_{1M}(0) + H_{1M}(1)z^{-1} \dots + H_{1M}(N)z^{-N} \\
H_{21}(0) + H_{21}(1)z^{-1} \dots + H_{21}(N)z^{-N} & H_{22}(0) + H_{22}(1)z^{-1} \dots + H_{22}(N)z^{-N} & \dots & H_{2M}(0) + H_{2M}(1)z^{-1} \dots + H_{2M}(N)z^{-N} \\
\vdots & \vdots & \ddots & \vdots \\
H_{M1}(0) + H_{M1}(1)z^{-1} \dots + H_{M1}(N)z^{-N} & H_{M2}(0) + H_{M2}(1)z^{-1} \dots + H_{M2}(N)z^{-N} & \dots & H_{MM}(0) + H_{MM}(1)z^{-1} \dots + H_{MM}(N)z^{-N}
\end{bmatrix}$$

$$= \begin{bmatrix}
H_{11}(0) & H_{11}(1) & \dots & H_{11}(N) & H_{12}(0) & H_{12}(1) & \dots & H_{12}(N) & \dots & H_{1M}(0) & H_{1M}(1) & \dots & H_{1M}(N) \\
H_{21}(0) & H_{21}(1) & \dots & H_{21}(N) & H_{22}(0) & H_{22}(1) & \dots & H_{22}(N) & \dots & H_{2M}(0) & H_{2M}(1) & \dots & H_{2M}(N) \\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
H_{M1}(0) & H_{M1}(1) & \dots & H_{M1}(N) & H_{M2}(0) & H_{M2}(1) & \dots & H_{M2}(N) & \dots & H_{MM}(0) & H_{MM}(1) & \dots & H_{MM}(N)
\end{bmatrix}
\begin{bmatrix}
1 & 0 & \dots & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 & \dots & 0 \\
z^{-1} & 0 & \dots & 0 & 1 & z^{-1} & \dots & 0 & 0 & z^{-1} & \dots & 0 & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
z^{-N} & 0 & \dots & 0 & 0 & 0 & \dots & 0 & z^{-N} & 0 & \dots & 0 & 0 \\
0 & 1 & \dots & 0 & 0 & 1 & \dots & 0 & 0 & 0 & \dots & 0 & 0 \\
0 & 0 & \dots & 0 & 0 & z^{-1} & \dots & 0 & 0 & 0 & \dots & 0 & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & \dots & 0 & 0 & z^{-N} & \dots & 0 & 0 & 0 & \dots & 0 & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & \dots & 0 & 0 & 0 & \dots & 1 & 0 & 0 & \dots & 1 & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & \dots & 0 & 0 & 0 & \dots & z^{-1} & 0 & 0 & \dots & z^{-1} & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & \dots & 0 & 0 & 0 & \dots & 0 & 0 & 0 & \dots & 0 & 0
\end{bmatrix}$$

The two dimensional transfer function is given by

$$G(z_1, z_2) = \sum_{k=0}^K \sum_{l=0}^K a_{k,l} z_1^{-k} z_2^{-l} \quad (10)$$

Notice that the elements $a_{k,l}$ can be arranged in matrix form as shown below

$$\begin{bmatrix} a_{00} & a_{01} & \cdots & a_{0K} \\ a_{10} & a_{11} & \cdots & a_{1K} \\ \vdots & \vdots & \ddots & \vdots \\ a_{K0} & a_{K1} & \cdots & a_{KK} \end{bmatrix} = \begin{bmatrix} \leftarrow a_0^T \rightarrow \\ \leftarrow a_1^T \rightarrow \\ \vdots \\ \leftarrow a_K^T \rightarrow \end{bmatrix} \quad (11)$$

Then the two dimensional transfer function can be written as

$$\begin{aligned} G(z_1, z_2) &= [a_{00} + a_{01}z_2^{-1} + \cdots + a_{0K}z_2^{-K}] \\ &+ z_1^{-1}[a_{10} + a_{11}z_2^{-1} + \cdots + a_{1K}z_2^{-K}] + \cdots \\ &+ z_1^{-K}[a_{K0} + a_{K1}z_2^{-1} + \cdots + a_{K2}z_2^{-K}] \end{aligned} \quad (12)$$

$$\left\{ \begin{bmatrix} a_{00} & a_{01} & \cdots & a_{0K} & a_{10} & a_{11} & \cdots & a_{1K} & \cdots & a_{K0} & a_{K1} & \cdots & a_{KK} \end{bmatrix} \begin{bmatrix} 1 & 0 & \cdots & 0 \\ z_2^{-1} & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ z_2^{-K} & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ 0 & z_2^{-1} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & z_2^{-K} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \\ 0 & 0 & \cdots & z_2^{-1} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & z_2^{-K} \end{bmatrix} \begin{bmatrix} 1 \\ z_1^{-1} \\ \vdots \\ z_1^{-K} \end{bmatrix} \right\} \quad (13)$$

Now let $a_i = H_i a_0$, then

$$G(z_1, z_2) = \{[a_0^T I \quad a_0^T H_1^T \quad \cdots \quad a_0^T H_K^T] Z_2\} Z_1$$

where Z_2 is the middle matrix of (13) and Z_1 is the rightmost vector of (13). Using this result $G(z_1, z_2) = a_0^T$ can be written as

$$a_0^T \begin{bmatrix} H_{11}(0) + \cdots + H_{11}(N)z_2^{-N} & H_{12}(0) + \cdots + H_{12}(N)z_2^{-N} & \cdots & H_{1M}(0) + \cdots + H_{1M}(N)z_2^{-N} \\ H_{21}(0) + \cdots + H_{21}(N)z_2^{-N} & H_{22}(0) + \cdots + H_{22}(N)z_2^{-N} & \cdots & H_{2M}(0) + \cdots + H_{2M}(N)z_2^{-N} \\ \vdots & \vdots & \ddots & \vdots \\ H_{M1}(0) + \cdots + H_{M1}(N)z_2^{-N} & H_{M2}(0) + \cdots + H_{M2}(N)z_2^{-N} & \cdots & H_{MM}(0) + \cdots + H_{MM}(N)z_2^{-N} \end{bmatrix} \begin{bmatrix} 1 \\ z_1^{-1} \\ \vdots \\ z_1^{-K} \end{bmatrix}$$

Therefore

$$G(z_1, z_2) = a_0^T H(z_2) Z_1, \quad (14)$$

and the power spectral density for the MAR process is given as

$$P_{AR}(z_1, z_2) = \frac{1}{|H(z_1, z_2)|^2} = |a_0^T H(z_2)^{-1} Z_1|^2, \quad (15)$$

where $a_0 = \Sigma^{-1} [\sigma \ 0 \ 0 \ \dots \ 0]^T$, σ is the 2-D prediction error variance, and $H(z_2) = I + \sum_{k=1}^p A^T(k) z_2^{-k}$. In a later section, the resulting MAR 2D spectrum will be compared to the spectral estimate produced by the taking the classical periodogram of the actual radar data.

4 Multidimensional Autoregressive Model Order

The *AIC* for real multidimensional autoregressive processes was first introduced by Akaike in [21] and is one of the model order indicators most cited. The *AIC* criterion for real MAR processes [2] is given by

$$AIC(p) = N \ln(\|\hat{\Sigma}\|) + 2J^2 p \quad (16)$$

where p is the model order, $\hat{\Sigma}$ is the p th order estimate of the prediction error covariance matrix, J is the number of channels, and $\|\cdot\|$ indicates the determinant operation. Notice that since (16) was developed for real data, the direct application to the multichannel airborne radar problem (with complex signals) is not readily apparent. The following development of the *AIC* is a synthesis of the various *AIC* development found in the literature. It has been included here under a unifying notation for extension the radar data case.

It is assumed that a $J \times 1$ -vector X can be represented by the following multidimensional autoregressive model:

$$X(n) = - \sum_{m=1}^p A(m) X(n-m) + U(n), \quad (17)$$

where $A(m)$ is a $J \times J$ matrix and $U(n)$ is a random $J \times 1$ vector with the following relations

$$\begin{aligned} E\{U(n)\} &= 0, \{0 \in \mathbb{C}^J\} \\ E\{U(n)X^T(n-m)\} &= 0, \{0 \in \mathbb{C}^{J \times J}\} \\ E\{U(n)U(m)^T\} &= \delta(n-m)\Sigma \end{aligned}$$

where E is the expectation operator, and Σ is a positive definite $J \times J$ matrix.

An important assumption in the development of the *AIC* is the Gaussianity of the least squares estimate, denoted by $\hat{A}(m)$, of $A(m)$,

- $E_\infty \{\sqrt{N}[\hat{A}_{i,j}(m) - A_{i,j}(m)]\} = 0, \ i, j = 1, \dots, J, \ m = 1, \dots, p.$
- $E_\infty \{N[\hat{A}_{i_1, j_1}(m) - A_{i_1, j_1}(m)][\hat{A}_{i_2, j_2}(n) - A_{i_2, j_2}(n)]\} = \Sigma_{i_1, i_2} R_{xx}^{-1}(m, j_1; n, j_2).$

where E_∞ is the limiting expectation. Now, consider the one-step prediction error when $\hat{A}(m)$ is applied to another independent realization of $X(n)$

$$D(n) = \sum_{m=1}^p \left(A(m) - \hat{A}(m) \right) X(n-m) + U(n).$$

The variance of this error with respect to an independent realization $X(n)$ is derived next. Let

$$\begin{aligned} D(n)D^T(n) &= \left(\sum_{m=1}^p [A(m) - \hat{A}(m)]X(n-m) + U(n) \right) \left(\sum_{l=1}^p [A(l) - \hat{A}(l)]X(n-l) + U(n) \right)^T \\ &= \sum_{m=1}^p \Delta A(m)X(n-m) \sum_{l=1}^p X(n-l)\Delta A^T(l) + \sum_{m=1}^p \Delta A(m)X(n-m)U^T(n) \\ &\quad + U(n) \sum_{l=1}^p X^T(n-l)\Delta A^T(l) + U(n)U^T(n). \end{aligned}$$

where $\Delta A(m) = A(m) - \hat{A}(m)$. Taking the expectation w.r.t. X results in the following

$$\begin{aligned} E\{D(n)D^T(n)\} &= \sum_{m=1}^p \sum_{l=1}^p \Delta A(m)E\{X(n-m)X^T(n-l)\}\Delta A^T(l) \\ &\quad + \sum_{m=1}^p \Delta A(m)E\{X(n-m)U^T(n)\} \\ &\quad + \sum_{l=1}^p E\{U(n)X^T(n-l)\}\Delta A^T(l) + E\{U(n)U^T(n)\} \\ &= \Sigma + \sum_{m=1}^p \sum_{l=1}^p \Delta A(m)E\{X(n-m)X^T(n-l)\}\Delta A^T(l). \end{aligned}$$

Notice that an element of the product of three matrices can be represented as follows:

$$\begin{aligned} (A(BC))_{ih} &= \sum_j A_{ij}(BC)_{jh} \\ &= \sum_j \sum_g A_{ij}B_{jg}C_{gh} \end{aligned}$$

Letting $A = A(m)$, $B = R$, and $C = A^T(m)$ then

$$(ARA^T)_{ih} = \sum_{j=1}^J \sum_{g=1}^J A_{i,j}(m)A_{h,g}(l)R_{j,g}$$

Therefore the i, h -th element of the second term of $E\{D(n)D^T(n)\}$ is

$$\sum_{m=1}^p \sum_{l=1}^p \sum_{j=1}^J \sum_{g=1}^J \Delta A_{i,j}(m)\Delta A_{h,g}(l)R_{xx}(m, j; l, g).$$

Taking the E_∞ of the above expression results in

$$\begin{aligned} E_\infty\{E\{D(n)D^T(n)\}\} &= \frac{1}{N} \sum_{m=1}^p \sum_{l=1}^p \sum_{j=1}^J \sum_{g=1}^J E_\infty\{N[\Delta A_{i,j}(m)][\Delta A_{h,g}(l)]\}R_{xx}(m, j; l, g) \\ &= \frac{1}{N} \Sigma_{i,h} \sum_{m=1}^p \sum_{l=1}^p \sum_{j=1}^J \sum_{g=1}^J R_{xx}^{-1}(m, j; l, g)R_{xx}(m, j; l, g). \end{aligned}$$

The above result can be thought of as taking the trace of an $Jp \times Jp$ identity matrix. Thus the above operation is equal to $\frac{Jp}{N}\Sigma$. Then, the variance of the prediction error is equal to

$$E_{\infty}\{E_x\{D(n)^T D(n)\}\} = \Sigma + \frac{Jp}{N}\Sigma. \quad (18)$$

The determinant of this expression is equal to

$$\|E_{\infty}\{E_x\{D(n)D(n)^T\}\}\| = \left(1 + \frac{Jp}{N}\right)^J \|\Sigma\|. \quad (19)$$

This formula is called the multiple final prediction error (MFPE). However, the approximation of the covariance matrix must be taken into account since the true Σ is often unavailable. The derivation of $\hat{\Sigma}$ follows as:

$$\begin{aligned} \hat{\Sigma}_p &= \frac{1}{N} \sum_{n=1}^N \left(X(n) - \hat{X}(n) \right) \left(X(n) - \hat{X}(n) \right)^T \\ &= \frac{1}{N} \sum_{n=1}^N - \left(\sum_{m=1}^p A(m)X(n-m) + U(n) + \sum_{m=1}^p \hat{A}(m)X(n-m) \right) \\ &\quad \left(- \sum_{l=1}^p A(l)X(n-l) + U(n) + \sum_{l=1}^p \hat{A}(l)X(n-l) \right)^T \\ &= \frac{1}{N} \sum_{n=1}^N \left\{ U(n) - \sum_{m=1}^p \Delta A(m)X(n-m) \right\} \left\{ U(n) - \sum_{l=1}^p \Delta A(l)X(n-l) \right\}^T \\ &= \frac{1}{N} \sum_{n=1}^N \left\{ U(n)U^T(n) - U(n) \sum_{l=1}^p X^T(n-l)\Delta A^T(l) - \sum_{m=1}^p \Delta A(m)X(n-m)U^T(n) \right. \\ &\quad \left. + \sum_{m=1}^p \Delta A(m)X(n-m) \sum_{l=1}^p X^T(n-l)\Delta A^T(l) \right\}. \end{aligned} \quad (20)$$

However, this simplifies by noting from the orthogonality principle :

$$\begin{aligned} 0 &= \sum_{n=1}^N \left(X(n) - \hat{X}(n) \right) X^T(n-l) \\ &= \sum_{n=1}^N \left(U(n) - \sum_{m=1}^p A(m)X(n-m) + \sum_{m=1}^p \hat{A}(m)X(n-m) \right) X^T(n-l) \\ &= \sum_{n=1}^N \left(U(n) - \sum_{m=1}^p \Delta A(m)X(n-m) \right) X^T(n-l). \end{aligned}$$

Thus

$$\frac{1}{N} \sum_{n=1}^N U(n)X^T(n-l) = \frac{1}{N} \sum_{n=1}^N \sum_{m=1}^p \Delta A(m)X(n-m)X^T(n-l).$$

Applying the above equality to the second and third term of the last line of equation (20) results in

$$\begin{aligned}
&= - \sum_{m=1}^p \Delta A(m) \left(U(n) X^T(n-m) \right)^T - \sum_{l=1}^p \left(U(n) X^T(n-l) \right) \Delta A^T(l) \\
&= - \sum_{m=1}^p \Delta A(m) \left[\sum_{l=1}^p \Delta A(l) X(n-l) X^T(n-m) \right]^T - \sum_{l=1}^p \left[\sum_{m=1}^p \Delta A(m) X(n-m) X^T(n-l) \right] \Delta A^T(l) \\
&= - \sum_{m=1}^p \sum_{l=1}^p \Delta A(m) X(n-m) X^T(n-l) \Delta A^T(l) - \sum_{m=1}^p \sum_{l=1}^p \Delta A(m) X(n-m) X^T(n-l) \Delta A^T(l) \\
&= -2 \sum_{m=1}^p \sum_{l=1}^p \Delta A(m) X(n-m) X^T(n-l) \Delta A^T(l).
\end{aligned}$$

Therefore

$$\begin{aligned}
\hat{\Sigma}_p &= \frac{1}{N} \sum_{n=1}^N U(n) U^T(n) \\
&\quad - 2 \sum_{m=1}^p \sum_{l=1}^p \Delta A(m) \frac{1}{N} \sum_{n=1}^N X(n-m) X^T(n-l) \Delta A^T(l) \\
&\quad + \sum_{l=1}^p \sum_{m=1}^p \Delta A(m) \frac{1}{N} \sum_{n=1}^N X(n-m) X^T(n-l) \Delta A(l) \\
&= \frac{1}{N} \sum_{n=1}^N U(n) U^T(n) \\
&\quad - \sum_{l=1}^p \sum_{m=1}^p \Delta A(m) \frac{1}{N} \sum_{n=1}^N X(n-m) X^T(n-l) \Delta A^T(l).
\end{aligned}$$

Then taking the infinity expectation results in

$$\hat{\Sigma}_p = \left(\Sigma - \frac{Jp}{N} \Sigma \right), \quad (21)$$

yielding

$$\|\Sigma\| = \left(1 - \frac{Jp}{N} \right)^{-J} \|\hat{\Sigma}_p\|. \quad (22)$$

Therefore putting (22) in (18) results in the proposed model order which gives the minimum of the multiple final prediction error

$$MFPE(p) = \left(1 + \frac{Jp}{N} \right)^J \left(1 - \frac{Jp}{N} \right)^{-J} \|\hat{\Sigma}_p\|. \quad (23)$$

The development of the multichannel *AIC* in (16) follows from the above result by finding the maximum likelihood of the joint Gaussian density function. It is assumed that a set of independent Gaussian real vectors with positive definite time-invariant covariance matrix Σ is given. The joint Gaussian density function is

$$f_{NJ}(U) = \frac{1}{(2\pi)^{\frac{1}{2}JN} \|\Sigma\|^{\frac{1}{2}N}} e^{-\frac{1}{2} \sum_{k=1}^N U^T(k) \Sigma^{-1} U(k)}. \quad (24)$$

Taking the natural log of the above function and multiplying it by -2 results in

$$\begin{aligned} L &= JN \ln(2\pi) + N \ln \|\Sigma\| + \sum_{n=1}^N U^T(n) \Sigma^{-1} U(n) \\ &= JN \ln(2\pi) + N \ln \|\Sigma\| + N \text{tr} \left(\frac{1}{N} \sum_{n=1}^N U^T(n) U(n) \Sigma^{-1} \right) \\ &= JN \ln(2\pi) + N \ln \|\Sigma\| + N \text{tr} (\Sigma_0 \Sigma^{-1}) \end{aligned} \quad (25)$$

where $\Sigma_0 = \frac{1}{N} \sum U^T(n) U(n)$. Taking the partial derivative with respect to Σ_{ij} of the above expression results in

$$\begin{aligned} \frac{\partial L}{\partial \Sigma_{ij}} &= N \frac{\partial}{\partial \Sigma_{ij}} \ln \|\Sigma\| + N \frac{\partial}{\partial \Sigma_{ij}} \text{tr} (\Sigma_0 \Sigma^{-1}) \\ &= N \text{tr} \left(\frac{\partial \Sigma}{\partial \Sigma_{ij}} \Sigma^{-1} \right) - N \text{tr} \left(\Sigma_0 \Sigma^{-1} \frac{\partial \Sigma}{\partial \Sigma_{ij}} \Sigma^{-1} \right). \end{aligned}$$

Equating this expression with zero results in

$$\text{tr} \left(\frac{\partial \Sigma}{\partial \Sigma_{ij}} \Sigma \right) = \text{tr} \left(\Sigma_0 \Sigma^{-1} \frac{\partial \Sigma}{\partial \Sigma_{ij}} \Sigma^{-1} \right).$$

Then $\Sigma_{ji}^{-1} = (\Sigma_0 \Sigma^{-1} \Sigma^{-1})_{ji}$, which implies that $\Sigma_0 = \Sigma$. Therefore, to maximize (25), $N \ln \|\Sigma_0\|$ needs to be maximized. Now substituting for Σ the MFPE estimate, (23) yields

$$\begin{aligned} N \ln \|\Sigma\| &= N \ln \left\| \left(1 + \frac{Jp}{N} \right) \left(1 - \frac{Jp}{N} \right)^{-1} \hat{\Sigma}_p \right\| \\ &= N \ln \left(1 + \frac{Jp}{N} \right)^J + N \ln \left(1 - \frac{Jp}{N} \right)^{-J} + N \ln \|\hat{\Sigma}_p\| \\ &= JN \ln \left(1 + \frac{Jp}{N} \right) - JN \ln \left(1 - \frac{Jp}{N} \right) + N \ln \|\hat{\Sigma}_p\| \\ &\equiv JN \frac{Jp}{N} + JN \frac{Jp}{N} + N \ln \|\hat{\Sigma}_p\| \\ &= 2J^2 p + N \ln(\|\hat{\Sigma}\|). \end{aligned}$$

This result is called the *AIC* and it has been shown that this result and (23) predict the same model order [4].

4.1 AIC limitations

It has been shown that the *AIC* criterion tends to select too high order models and will overestimate the true order of a finite order autoregressive model [22]. Moreover, Nuttall presented in [5] an upper bound on the model order for which the *AIC* may be considered to be reliable. This upper bound is given by

$$\max(p_{AIC}) \leq \frac{3\sqrt{N}}{J}. \quad (26)$$

The *AIC* should not be considered reliable for model orders exceeding $\max(p_{AIC})$. That is, while the *AIC* may indeed be computed for high model orders, it may be an inaccurate measure for the model order. Nuttall found this result by extending to the multidimensional case the upper bound for the real-valued univariate case presented by Akaike [5]. Basically, this upper bound results from taking the ratio of the number of scalar coefficients and the number of available scalar data points

$$\frac{J^2 p}{JN} \leq \frac{3}{\sqrt{N}}$$

solving for p in this equation and calling it p_{AIC} results in (26). Also notice that this upper bound is a tighter upper bound than that suggested by Akaike in [21], which is given by either $\frac{N}{10J}$ or $\frac{N}{5J}$.

For the present airborne radar case, a typical return data length is short and complex, thus limiting the usefulness of the *AIC* when considering Nuttall upper bound. This lead us to consider a justification for model order selection based on the prediction error power. In the next section we first study the relation between equation (26), which depends on the data length, and the prediction error power which is calculated for independent radar return realizations. These experimental simulations show that for large model order there is a reduction in prediction error power for the range cell under study but at the expense of parameter variance. This, in turn, leads to large prediction error power for independent range cells and thus could serve as an indicator for model order selection. The simulations also verify the Nuttall upper bound from the prediction error power view point.

5 Computer Simulations

In this section computer simulations are presented for the prediction error power, model order selection, prediction error power for independent data realizations, and spectral analysis of modeled as well as actual radar data from the Multi-Channel Airborne Radar Measurement (MCARM) project. The simulated return data were generated by MATLAB routines developed by Scientific Studies Corporation [23]. The parameters used for the synthesized radar data are given in the table below

5.1 Prediction Error Power for simulated data

Figures 4 5, 6, and 7 show the prediction error power for simulated airborne return data sets of lengths 64, 128, 200, and 550. These plots are generated as described in the flowchart shown below. The simulations

| | |
|-------------------------|--|
| $J = 14$ | Number of linear array elements |
| $N = 64, 128, 200, 550$ | Number of pulses in one (CPI) |
| $\phi_0 = 30$ | Array main beam azimuth angle (deg) |
| $f_{PRF} = 300$ | Pulse repetition frequency (Hz) |
| $f_C = 450$ | Transmit frequency (MHz) |
| <i>UNIFORM</i> | Array pattern |
| $H_p = 9$ | Platform altitude (km) |
| $V_p = 25$ | Platform velocity (m/s) |
| $r_c = 130$ | Range to desired ground clutter ring (km) |
| $\gamma = 20$ | Aircraft platform crab angle (deg) |
| $N_c = 361$ | Number of ground clutter patches in the clutter ring |
| $varn = 1$ | Noise power in each channel |
| $SNR = 3$ | Target signal-to-noise ratio (dB) |
| f_s | Spatial Frequency $\{-0.25\}$ |
| f_t | Temporal Frequency $\{0.25\}$ |
| $Nr = 10$ | Number of independent realizations |

Table 1: Simulation Parameters

are performed for MAR model orders from 1 through 10. Once the model order is selected the MAR parameters are calculated as described in (7) where biased estimates of the correlation matrices are used from the received signal. The prediction error power is calculated by taking the trace of the driving noise covariance matrix as calculated in (8). One salient feature of these prediction error power plots are that the difference between the PEP of order 1 and those of higher order are within 2.5dB. This means that any increase in the model order does not contribute significantly to any additional prediction error reduction. Also notice that since the prediction error power is calculated for a given model order the length of these plots is independent of the length of data under study.

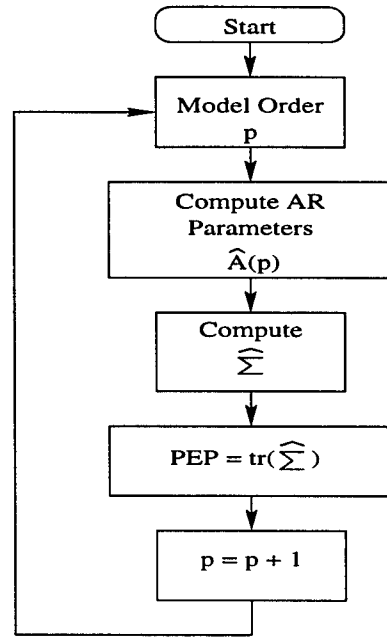


Figure 3: PEP flowchart

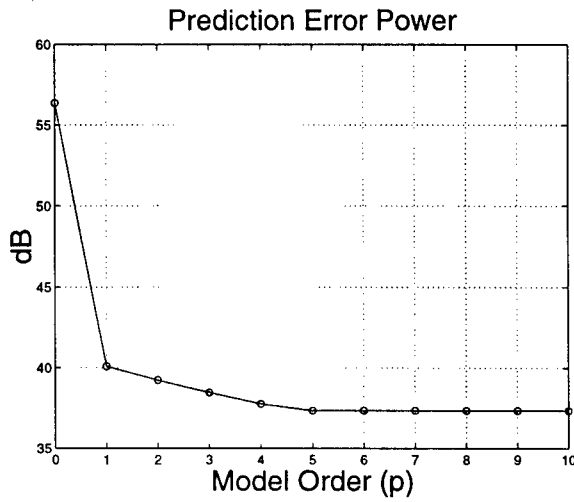


Figure 4: PEP for 64 data points

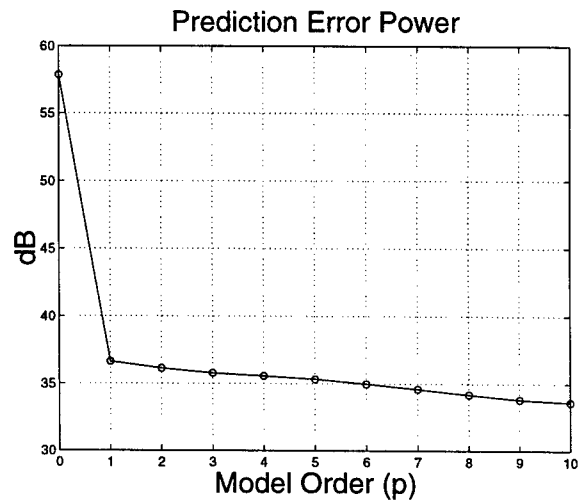


Figure 5: PEP for 128 data points

5.2 AIC for simulated data

Figures 8, 9, 10, and 11 show *AIC* plots for simulated data of lengths 64, 128, 200, and 550. The *AIC* is calculated as in (16), repeated here for clarity

$$AIC(p) = N \ln(\|\hat{\Sigma}\|) + 2J^2p$$

In general, the proper model order is usually that which minimizes the *AIC*. However, notice that these figures do not show the expected concave behavior for the *AIC* curves. These *AIC* curves are thus better

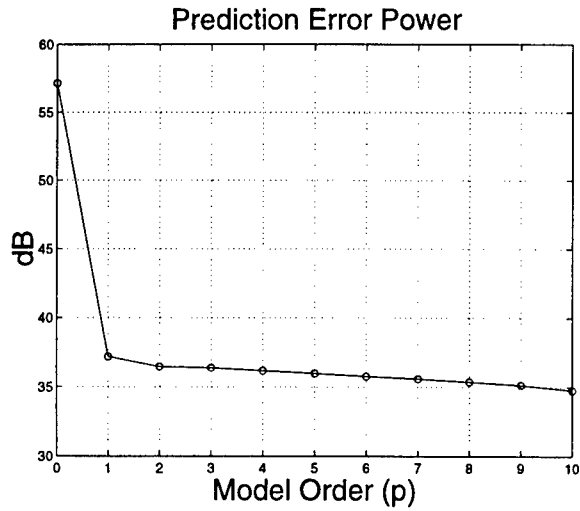


Figure 6: PEP for 200 data points

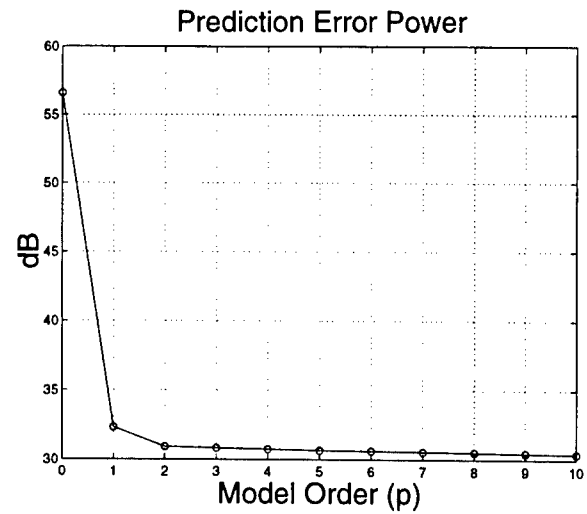


Figure 7: PEP for 550 data points

explained in conjunction with the Nuttall upper bound. Notice that for Figure 8 the minimum AIC is 8, while on the contrary the Nuttall upper bound for this data set dictates that it not exceed 1.7 as indicated in the table below. This indicates that, the AIC should be used cautiously for such short data sets.

| Data points | Nuttall upper bound | Model order estimate by AIC |
|-------------|---------------------|-----------------------------|
| 64 | 1.7 | 8 |
| 128 | 2.4 | 10 |
| 200 | 3.0 | 2 |
| 550 | 5.0 | 4 |

Table 2: Nuttall upper bound and AIC relation

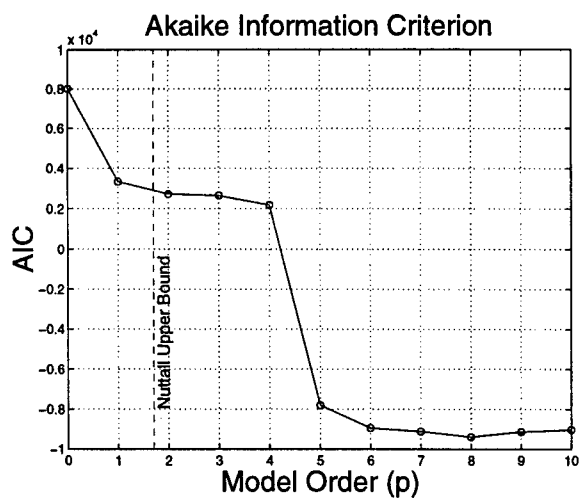


Figure 8: AIC for 64 data points

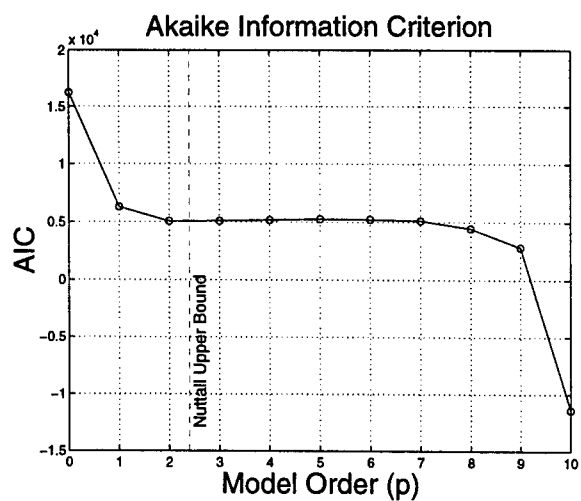


Figure 9: AIC for 128 data points

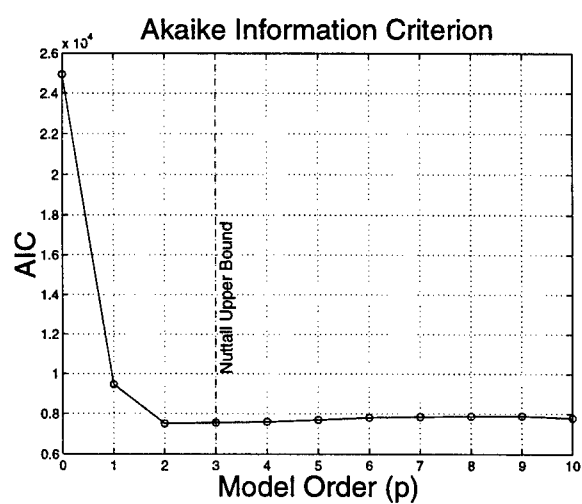


Figure 10: AIC for 200 data points

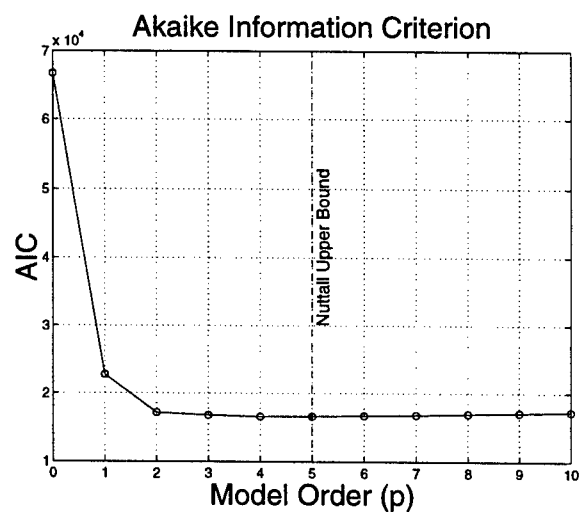


Figure 11: AIC for 550 data points

5.3 PEP for Independent Realizations

As discussed in the previous section the *AIC* has some limitations, in particular the short temporal support of the radar return signal restricts the use of the *AIC*. Therefore, an alternative means of justifying the MAR model order based on the prediction error power follows.

Model order selection information can be extracted from the prediction error power (PEP) of independent data realizations as shown in the flowchart below. The independent realizations, for a given data length, are generated with MATLAB routines supplied by [23]. For the case of actual radar return data, independence is related to different range cells for a particular data acquisition.

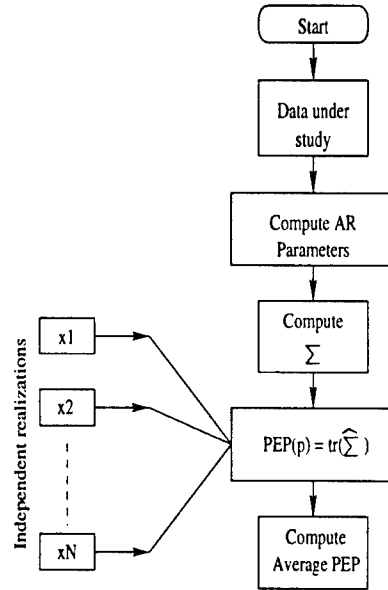


Figure 12: PEP flowchart for independent data realizations

The simulation consists in calculating the MAR parameter matrices for one range cell using (7) with biased estimates of the correlation matrices and using these parameters to calculate the prediction error power for independent data sets. Then, the prediction error power is calculated by taking the trace of the driving noise covariance matrix as calculated in (8).

The idea behind this simulation is to observe that for low model orders, the set of MAR parameters calculated for the range cell under study can be used for independent range cells. This observation can be seen in Figures 13, 14, 15, and 16. These plots show that indeed the average PEP (x - solid line) for a set of independent data realizations (dashed lines) varies dramatically, away from the PEP of the data under study (o - solid line), from the *AIC* results above the Nuttall upper bound. This relates that fact that the *AIC* should not be considered reliable for model orders exceeding the Nuttall upper bound. This indicates that the estimated MAR parameters for one data set applied to an independent data set from the same process have a relatively small variance for model orders less than the Nuttall limit. Also notice that as the data length increases the averaged prediction error power decreases since the MAR parameter variance decreases.

Finally we have that for short data sets, prediction error power analysis in conjunction with the Nuttall upper bound provide a crude indicator for the model order to use. In the next section a similar analysis and result is presented for actual return data sets.

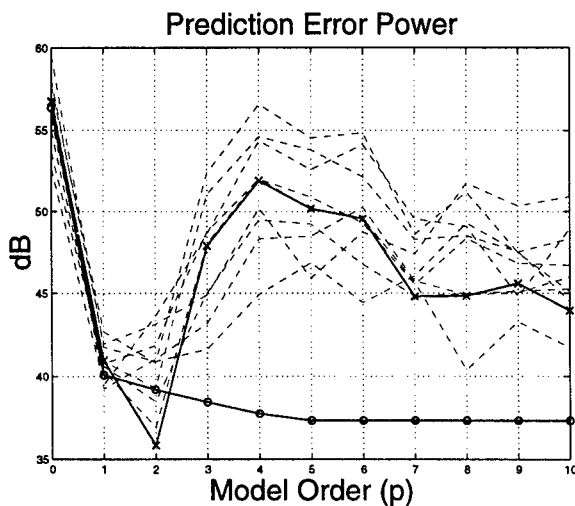


Figure 13: Independent realizations for 64 data points

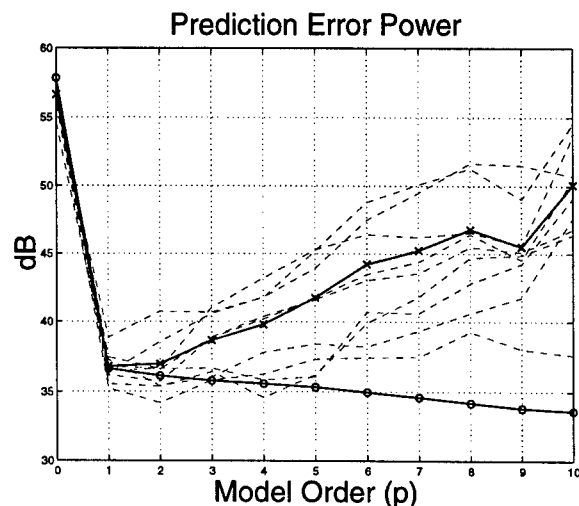


Figure 14: Independent realizations for 128 data points

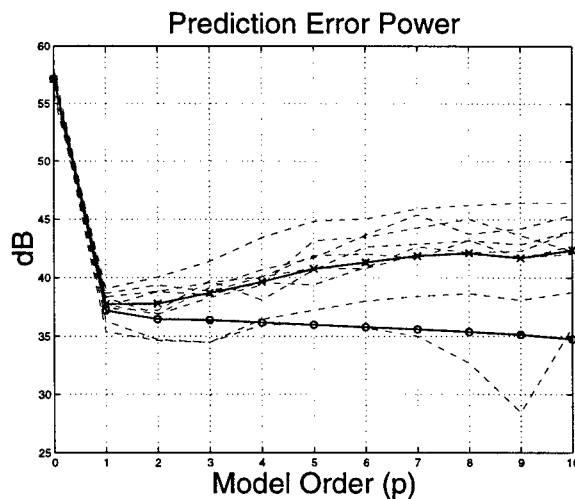


Figure 15: Independent realizations for 200 data points

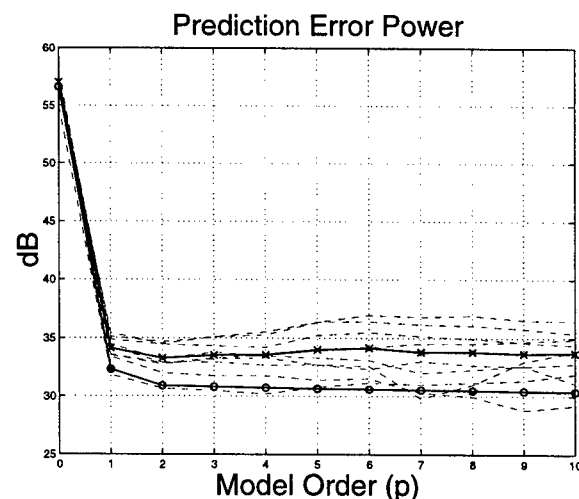


Figure 16: Independent realizations for 550 data points

5.4 MCARM data

This section presents simulations using Multi-Channel Airborne Radar Measurement (MCARM) data. A typical *AIC* curve for this type of data is shown in Figure 17. Notice that this curve does not show the expected concave behavior since the data has relatively short time support (128 time samples and 14 channels).

The figure also shows the Nuttall upper bound.

Figures 33 through 47 are typical prediction error plots for different range cells. Notice that, as expected, in all of these cases the prediction error power decreases with respect to the model order. This does not mean that the prediction error power yields the proper MAR model order. Also notice that in some cases the prediction error power decreases from about 2dB (Figure 33) to about 10dB (Figure 38). This might indicate that using a high model order will lead to a more accurate prediction of the data. However, as indicated by the upper bound *AIC* analysis, the upper limit for the model order order should not exceed 1.7 for the time support considered in these simulations. Moreover, if higher model orders are considered then we have an increase in variance which for the MAR model leads to undesired spurious peaks in the spectrum. This is reflected in the 2D-spectral plots 20 through 31 shown below for model order 2 and 10. These 2D-spectral plots were computed as described in section 3.

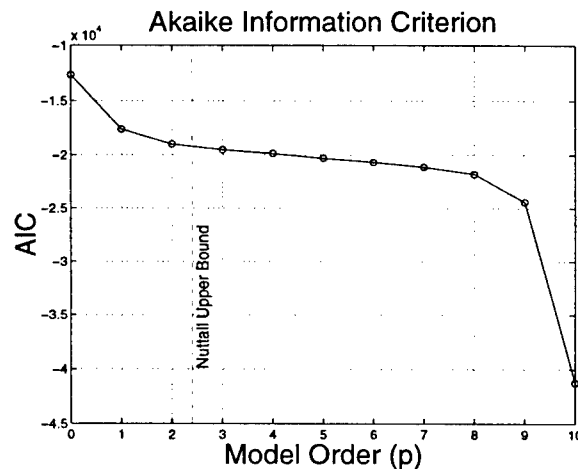


Figure 17: Typical AIC for MCARM data

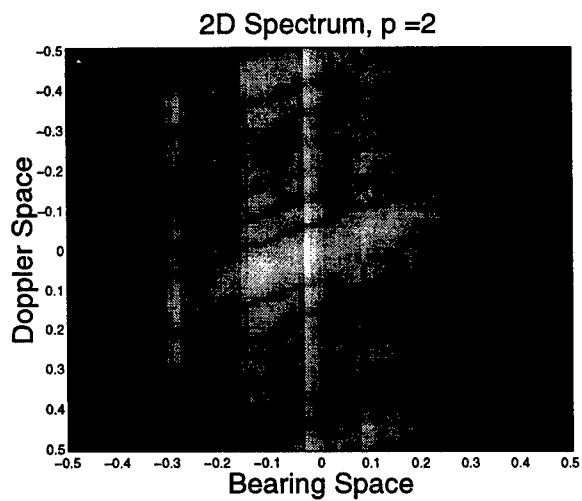


Figure 18: 2nd order MAR for acquisition #465, range cell 200

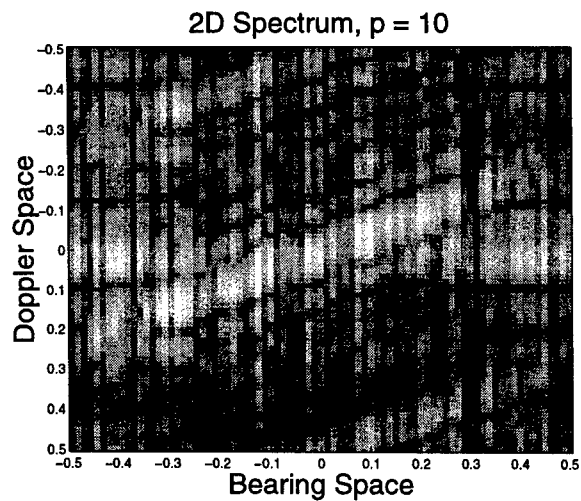


Figure 19: 10th order MAR for acquisition #465, range cell 200

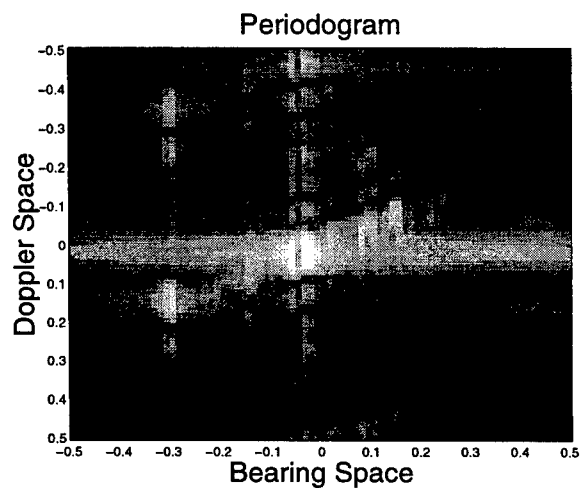


Figure 20: Periodogram for acquisition #465, range cell 200

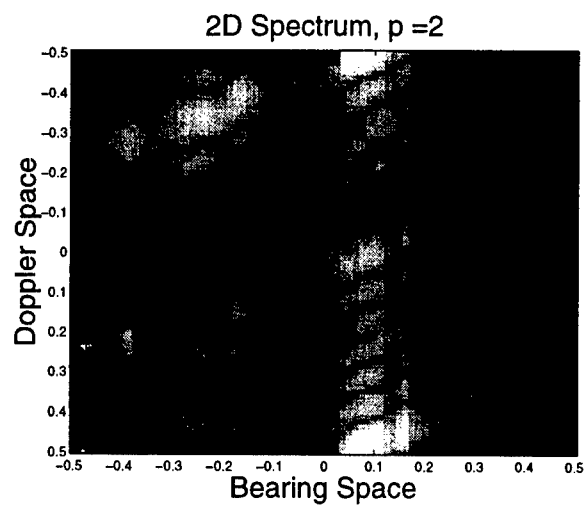


Figure 21: 2nd order MAR for acquisition #520, range cell 200

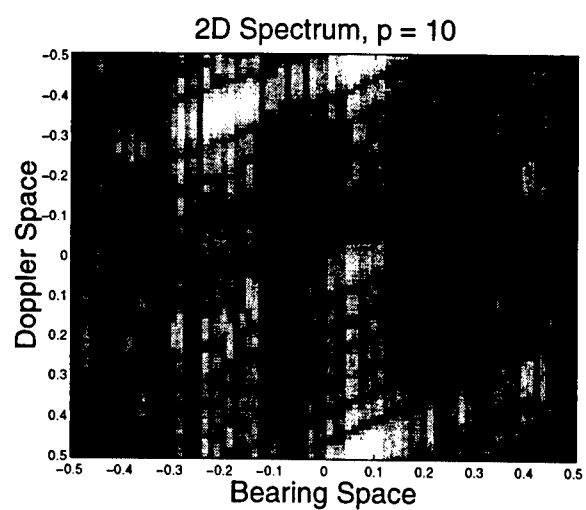


Figure 22: 10th order MAR for acquisition #520, range cell 200

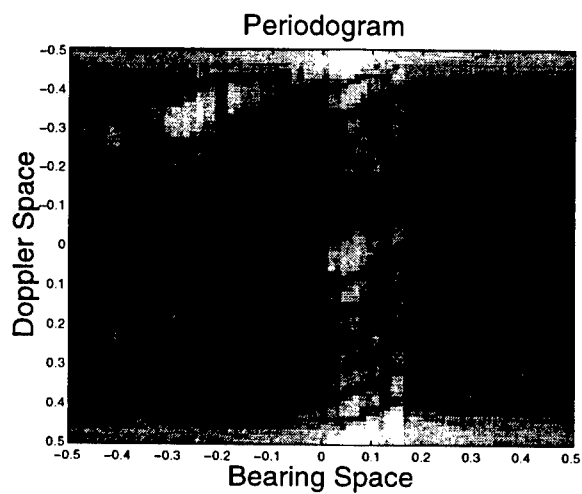


Figure 23: Periodogram for acquisition #520, range cell 200

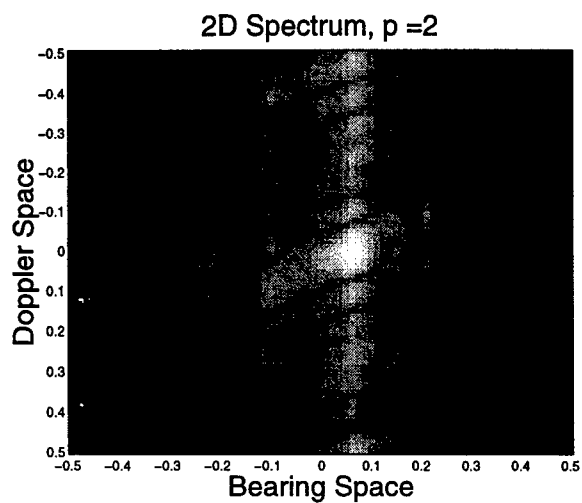


Figure 24: 2nd order MAR for acquisition #575, range cell 200

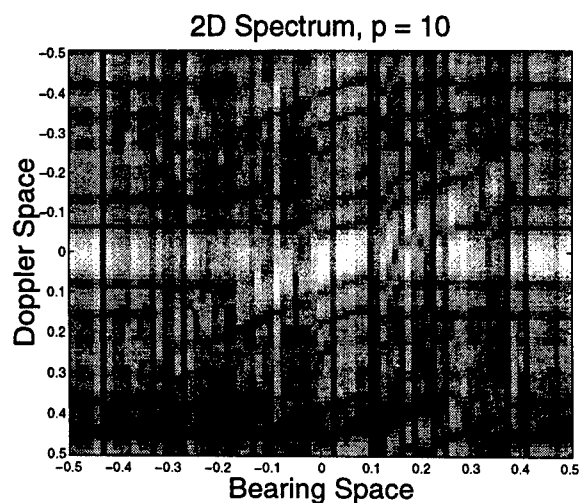


Figure 25: 10th order MAR for acquisition #575, range cell 200

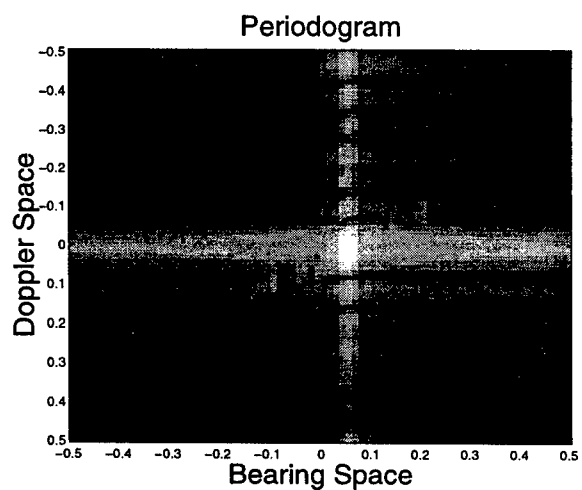


Figure 26: Periodogram for acquisition #575, range cell 200

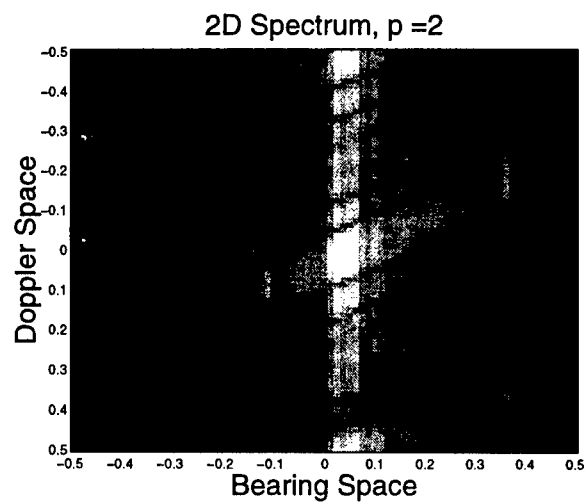


Figure 27: 2nd order MAR for acquisition #628, range cell 200

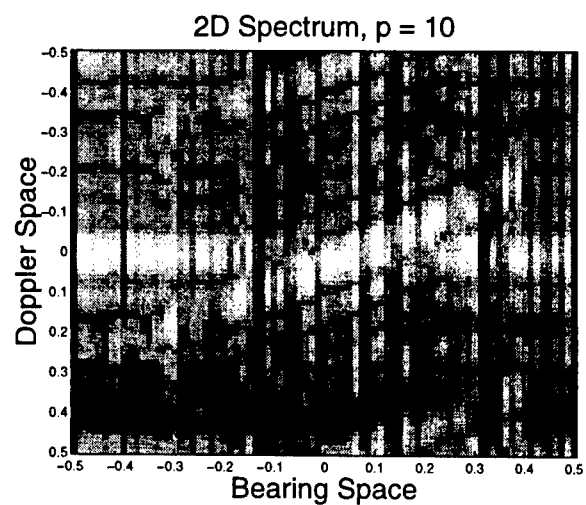


Figure 28: 10th order MAR for acquisition #628, range cell 200

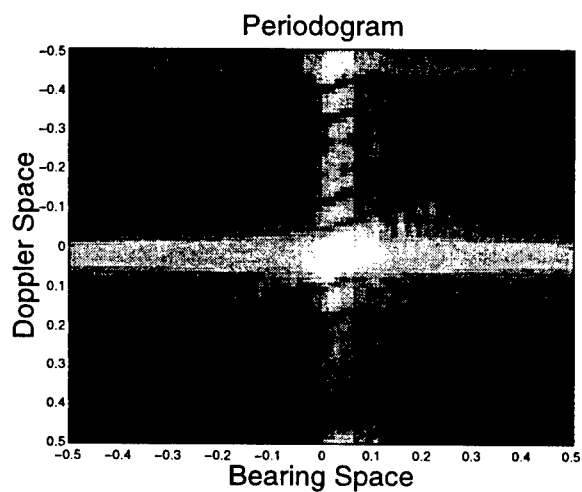


Figure 29: Periodogram for acquisition #628, range cell 200

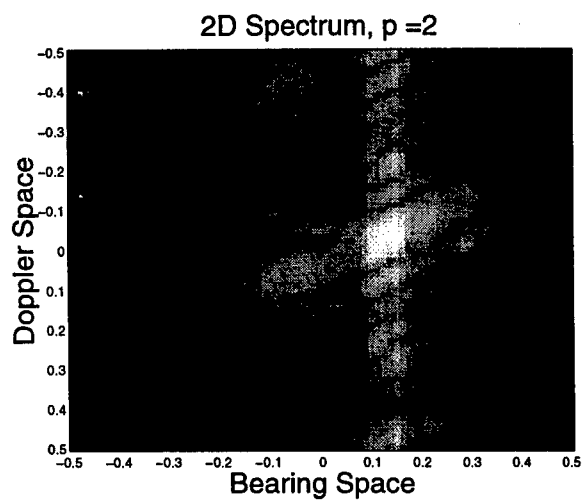


Figure 30: 2nd order MAR for acquisition #644, range cell 200

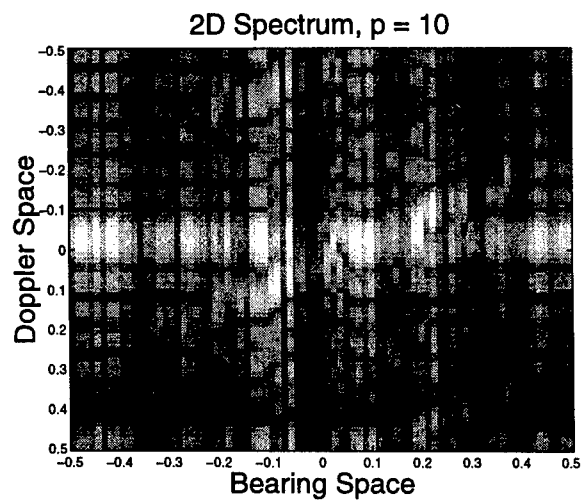


Figure 31: 10th order MAR for acquisition #644, range cell 200

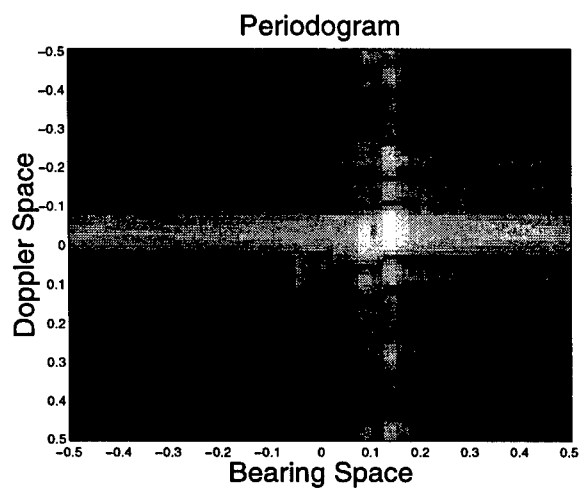


Figure 32: Periodogram for acquisition #644, range cell 200

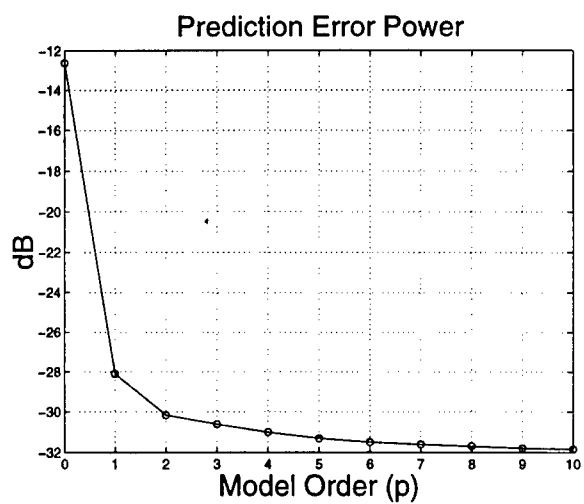


Figure 33: PEP (acquisition #465, range cell 105)

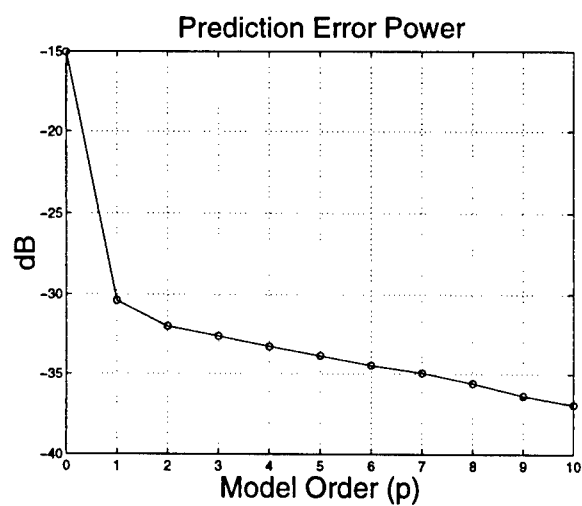


Figure 34: PEP (acquisition #465, range cell 200)

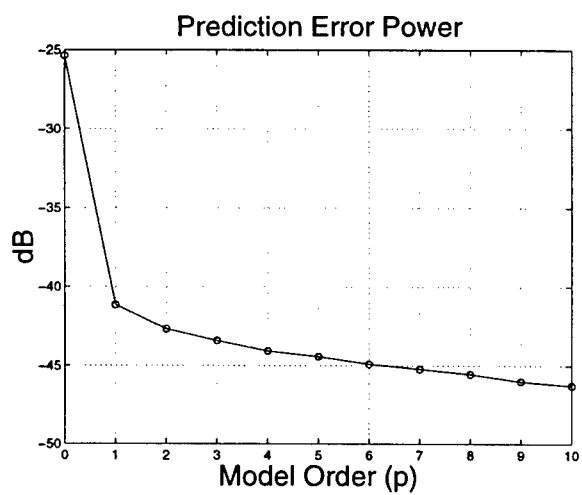


Figure 35: PEP (acquisition #465, range cell 350)

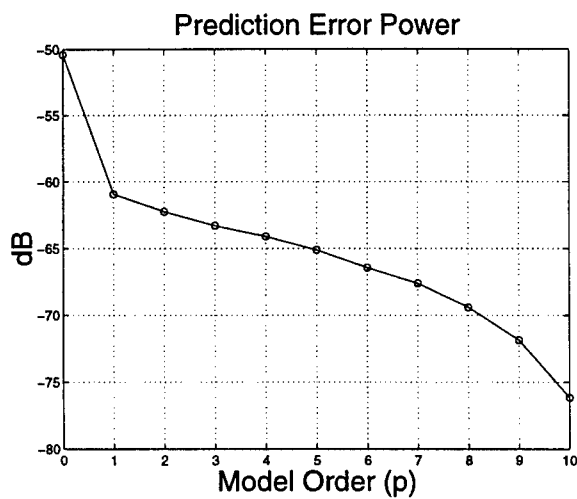


Figure 36: PEP (acquisition #520, range cell 105)

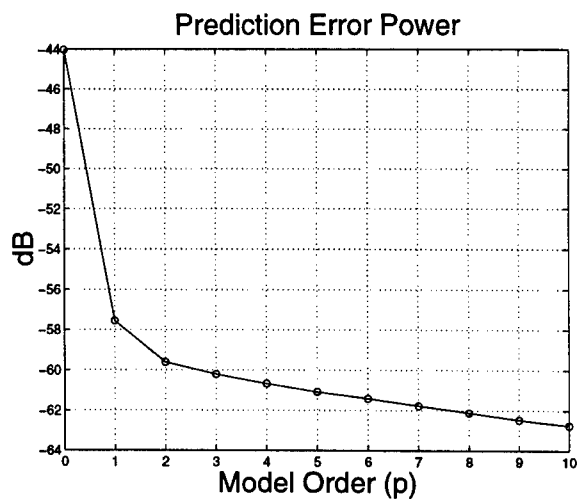


Figure 37: PEP (acquisition #520, range cell 200)

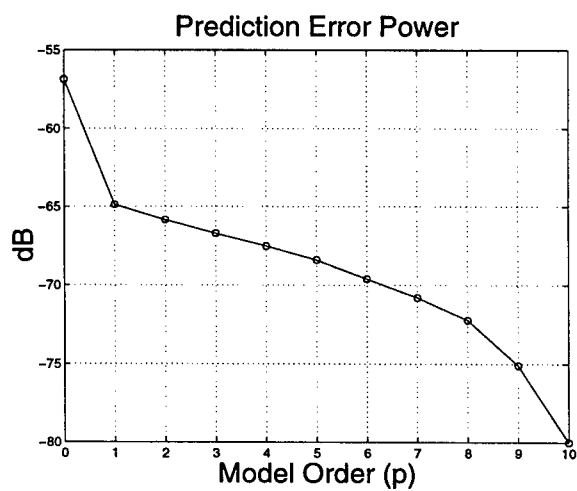


Figure 38: PEP (acquisition #520, range cell 350)

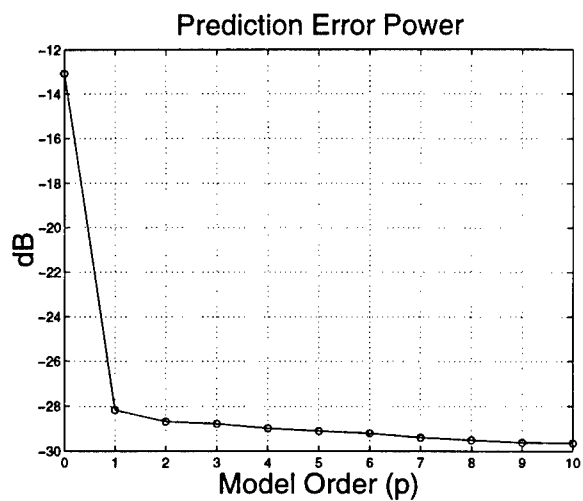


Figure 39: PEP (acquisition #575, range cell 105)

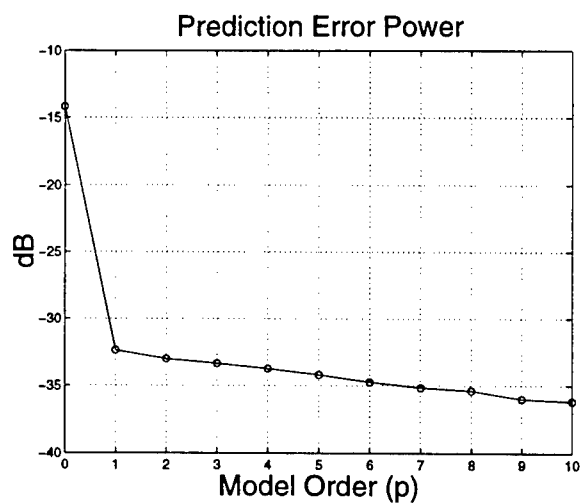


Figure 40: PEP (acquisition #575, range cell 200)

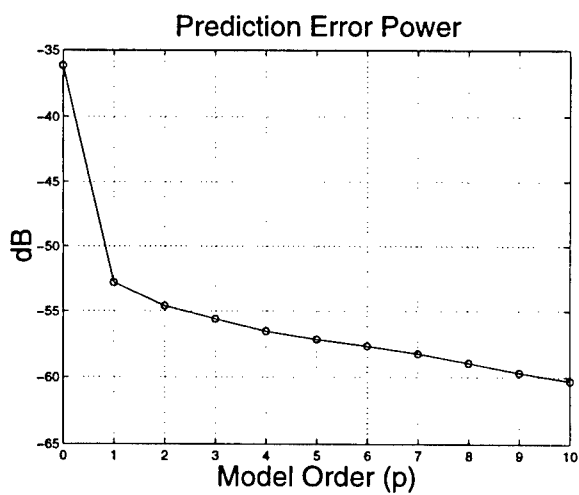


Figure 41: PEP (acquisition #575, range cell 350)

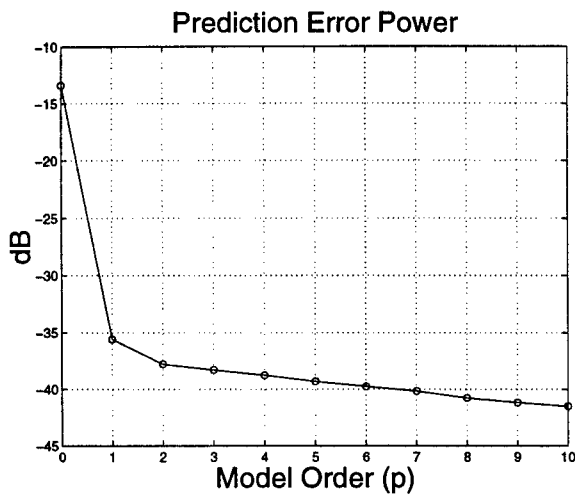


Figure 42: PEP (acquisition #628, range cell 105)

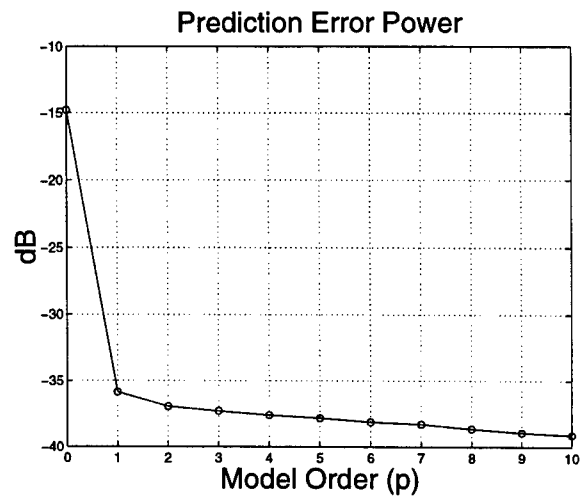


Figure 43: PEP (acquisition #628, range cell 200)

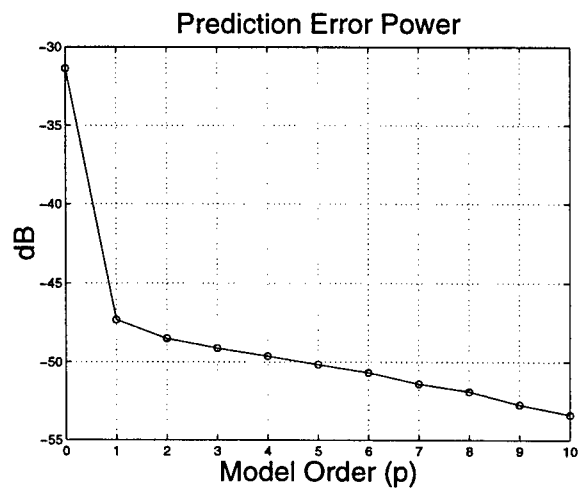


Figure 44: PEP (acquisition #628, range cell 350)

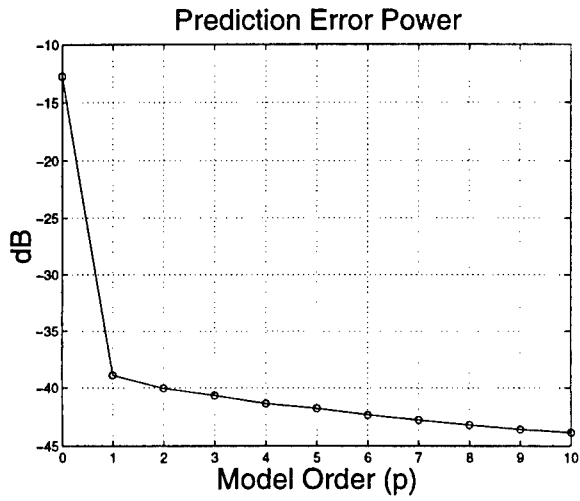


Figure 45: PEP (acquisition #644, range cell 105)

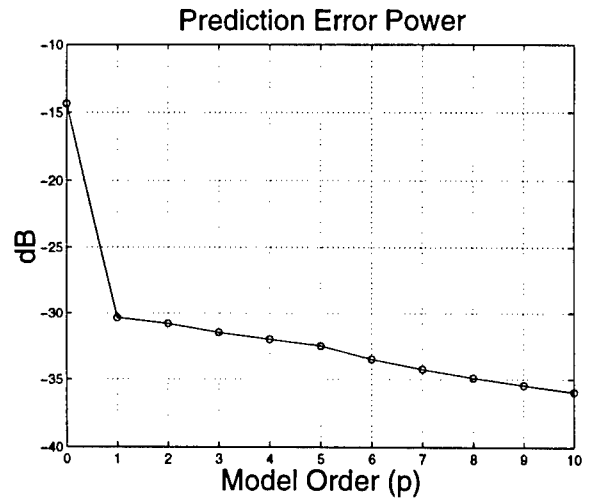


Figure 46: PEP (acquisition #644, range cell 200)

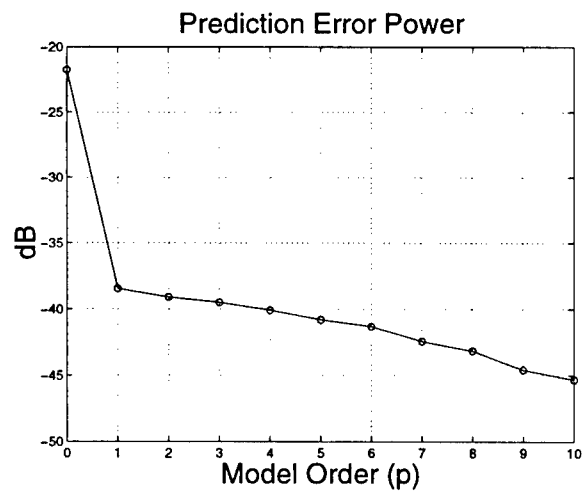


Figure 47: PEP (acquisition #644, range cell 350)

5.5 MCARM PEP for Independent Realizations

The following plots were generated as described by Figure 12 in section 5.3, where the independent data realizations are related to different range cells for a particular data acquisition. Notice that for this data, obtaining independent realizations of the process is impossible. However, if one assumes that different range cell data comes from a similar process, then the use of different range cell data sets may approximate the independent realization assumption.

Recall that the idea behind this experiment is to verify that the average PEP for a set of independent data realizations varies substantially (away from the PEP of the range cell under study) from the *AIC* results above the Nuttall upper bound. This suggests that the MAR parameters for one range cell applied to an independent range cell data set, from the similar process, have a relatively small variance for model orders less than the Nuttall limit. This type of analysis provides a crude indicator for the model order to use.

The data under study (o - solid line) is that of range cell 200 for the different acquisitions and the average prediction error power (x - solid line) was taken over 9 independent (different range cells) realizations (128 time samples and 14 channels). In particular, the independent realizations (dashed lines) are those of range cells 120, 140, 160, 180, 220, 240, 260, 280, and 300. Notice that all the plots show similar behavior as that of the simulated data. That is the set of independent realizations varies dramatically from the average PEP after a relatively small model order, except for acquisition #520 which does not show this general behavior. This might be attributed to the terrain for which this data was obtained. However, in general, we can assert from these simulations that a low model order MAR process can be used for the MCARM data set.

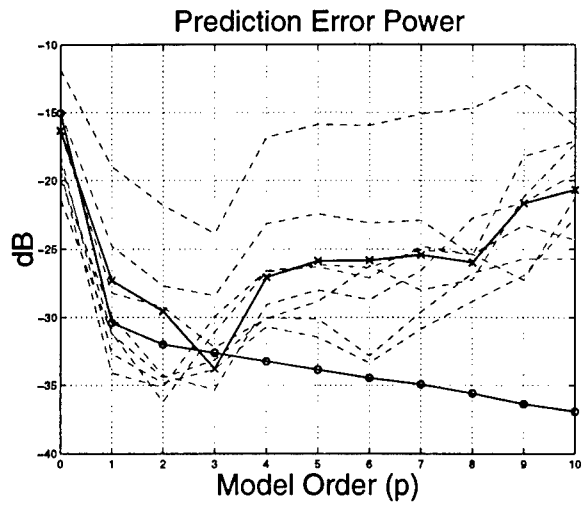


Figure 48: Independent realizations for acquisition #465

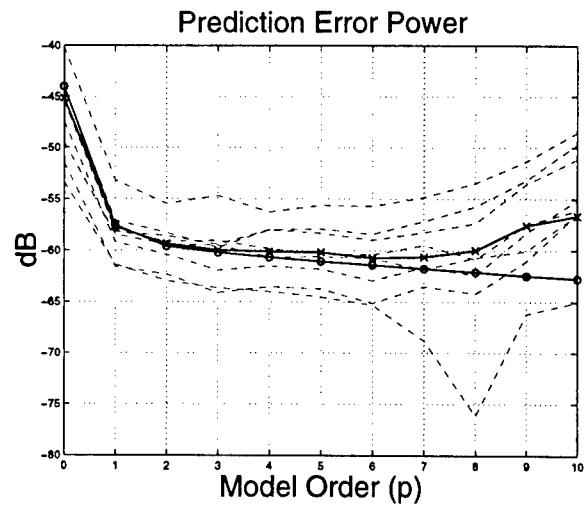


Figure 49: Independent realizations for acquisition #520

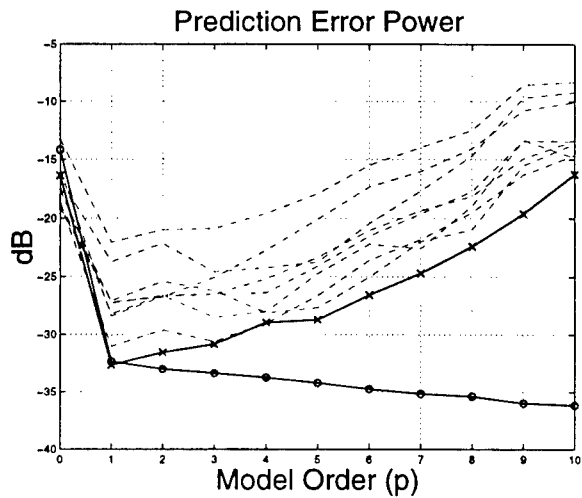


Figure 50: Independent realizations for acquisition #575

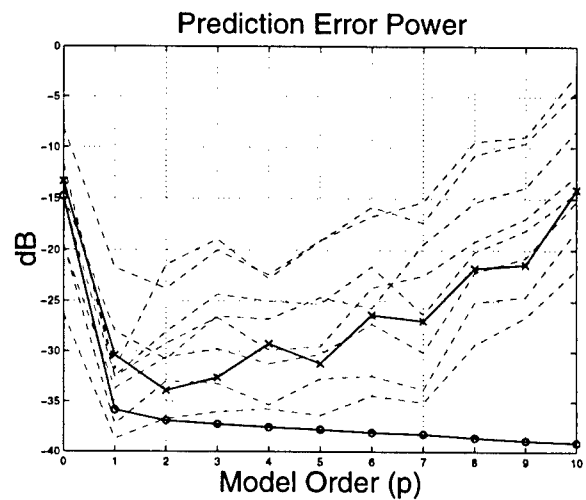


Figure 51: Independent realizations for acquisition #628

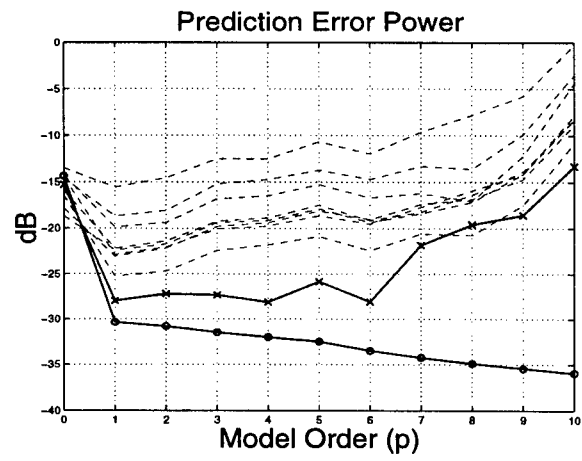


Figure 52: Independent realizations for acquisition #644

6 Conclusion

This work has focussed on the model order selection problem for airborne radar data in a innovations based detection setting. For proper detection processing, it is imperative that the signal model, here a multichannel autoregressive process (MAR), is capable of capturing the statistical characteristics of the clutter. To this end, the model order (p) must be large enough. However, the choice of a model order that is *too* large will be detrimental to detection processing due to the associated large parameter variance. To provide proper model order selection we have looked at two (somewhat related) criteria, the Akaike Information Criterion (AIC) and prediction error power. These criteria were investigated using both a physical radar signal model from SSC, and actual MCARM radar measurements.

The simulations using independent realization assumptions indicate that a MAR process with low model order ($p = 2$ or $p = 3$) might best be used for innovations based detection. The independent realizations simulations also verify the Nuttall upper bound for AIC confidence from the prediction error power view point. Results based on simulated data (which allow independent data realizations) indicate that, while large model orders achieve a reduction in prediction error power, this comes at the expense of excessive parameter variance. In turn, this leads to a large prediction error power for independent data sets serving as an indication of a too high MAR model order. In general, for the MCARM data sets, we have that the average PEP for a set of independent realizations varies substantially after a relatively small model order.

The study of the development of the standard AIC indicates that this model estimator should be used with care for radar return data. That is, the common assumptions in derivation of the AIC included: large data sets, real (i.e. non-complex) data, and Gaussianly distributed. These assumptions do not typically hold for actual radar returns, which is complex and has a short temporal support.

The MAR 2D-spectrum simulations included in this study show that a relative high model order results in spurious peaks while a low model order ($p = 2$) is capable of describing the clutter (from a visual comparison with the periodogram) in agreement with the PEP results. These results imply that a *low order* MAR process in an innovations based detection architectures may be used for target detection.

References

- [1] S. Haykin and A. Steinhardt, *Adaptive Radar Detection and Estimation*, John Wiley & Sons, Inc.: New York, 1992.
- [2] S. M. Kay, *Modern Spectral Estimation: Theory & Application*, New Jersey: Prentice-Hall, Inc., 1988.
- [3] S. Lawrence Marple, Jr., *Digital Spectral Analysis with Applications*, New Jersey: Prentice-Hall, Inc., 1987.
- [4] Richard H. Jones, "Identification and Autoregressive Spectrum Estimation," *IEEE Trans. Autom. Control*, Vol. AC-19, No. 6, pp. 894-897, Dec 1974.
- [5] A. H. Nuttall, "Multivariate Linear Predictive Spectral Analysis Employing Weighted Forward and Backward Averaging: A Generalization of Burg's Algorithm," *Naval Underwater System Center, New London, CT. Tech. Rep. 5501*, Oct. 13, 1976.
- [6] J. H. Michels, "Multichannel Detection Using The Discrete-Time Model-Based Innovations Approach," *Ph.D Dissertation*, Syracuse University, New York, May 1991.
- [7] J. P. LeBlanc, "Multichannel Autoregressive Modeling and Spectral Estimation Methods for Airborne Radar Environment," *AFOSR, Final Report*, pp. 16-1-16-19, Aug. 1996.
- [8] J. Castro and J. LeBlanc, "Model Order Selection For Multidimensional Innovations Based Detection in Airborne Radar," *IEEE Radar Conference*, Dallas, Texas, May 12-13, 1998 (to appear).
- [9] Space-Time Adaptive Processing (STAP) in Airborne Radar Applications, "J. Michels, T. Tsao, B. Himed, and M. Rangaswamy," *International Conference Signal Processing and Communications*, Canary Islands, Spain, February 11-14, 1998.
- [10] H. Akaike, "Information Theory and an Extension of the Maximum Likelihood Principle," *2nd International Symposium on Information Theory*, Armenia, USSR, September 2-8, 1971.
- [11] H. L. Van Trees, *Detection, Estimation, and Modulation Theory: Part I, II*, New York: John Wiley & Sons, Inc., 1968.
- [12] P. A. S. Metford, S. Haykin, and D. P. Taylor, "An Innovations Approach to Discrete-Time Detection Theory," *IEEE Trans. Inform. Theory*, Vol. IT-28, No. 2, March 1982.
- [13] P. A. S. Metford, S. Haykin, "Experimental Analysis of an Innovations-Based detection algorithm for Surveillance Radar," *IEE Proc.*, Vol. 132, Pt. F, No. 1, Feb 1985.
- [14] L. E. Brennan and I. S. Reed, "Theory of adaptive radar," *IEEE Tras. Aerospace Electron. Sys.* , Vol. AES-9, No. 2, March 1973.
- [15] B. Widrow and S. D. Stearns, *Adaptive Signal Processing*, New Jersey: Prentice-Hall, Inc., 1985.
- [16] Jaime R. Román, Dennis W. Davis, and James H. Michels, "Multichannel Parametric Models for Airborne Phased Array Clutter," *NATRAD 1997*, Syracuse, NY, May 1997.
- [17] M. Aoki, *State Space Modeling of Time Series*, New York: Springer-Verlag, 1990.

- [18] K. Ogata, *Discrete-Time Control Systems*, New Jersey: Prentice-Hall, Inc., 1987.
- [19] S. Haykin, *Adaptive Filter Theory*, New Jersey: Prentice-Hall, Inc., 1991.
- [20] C. W. Therrien, "Relations Between 2-D and Multichannel Linear Prediction," *IEEE Trans. on Acoust. Speech and Signal Proc.*, Vol. ASSP-29, No. 3, June, 1981.
- [21] H. Akaike, "Autoregressive Model Fitting for Control," *Ann. Inst. Statist. Math.*, Vol. 23, pp. 163-180, 1971.
- [22] W. A. Fuller, *Introduction to Statistical Time Series*, 2ed. pp. 438-439, New York: John Wiley & Sons, Inc., 1996.
- [23] J. R. Roman, D. W. Davis, "Multichannel System Identification using Output Data Techniques Vol II Final Report No. SSC-TR-96-02," *Scientific Studies Corp., Palm Beach Gardens, FL*, October, 1996.
- [24] H. Mann and A. Wald, "On the statistical treatment of linear stochastic difference equations," *Econometrica*, Vol. 11, pp. 173-220, 1943.

Related references

- [25] H. Akaike, "Maximum likelihood identification of Gaussian autoregressive moving average models," *Biometrika*, Vol. 60, No. 2, pp. 255-265, 1973.
- [26] H. Akaike, "Fitting autoregressions for prediction," *Ann. Inst. Statist. Math.*, Vol. 21, pp. 243-247, 1969.
- [27] D. A. S. Fraser, *Nonparametric Methods in Statistics*, New York: John Wiley & Sons, Inc., 1957.
- [28] P. R. Halmos and L. J. Savage, "Application of the Radon-Nykodym Theorem to the Theory of Sufficient Statistics," *Ann. Math. Statist.*, Vol. 20, 1949, pp. 225-241.
- [29] S. Haykin, J. Litva, and T. J. Shepherd (Eds.), *Radar Array Processing*, New York: Springer-Verlag, 1993.
- [30] D. Kazakos and P. P. Kazakos, *Detection and Estimation*, New York: Computer Science Press, 1990.
- [31] S. Kullback, *Information Theory and Statistics*, New York: John Wiley & Sons, Inc., 1959.
- [32] S. Kullback and R. A. Leibler, "On Information and Sufficiency," *Ann. Math. Statist.*, Vol. 22, 1951, pp. 79-86.
- [33] H. O. Lancaster, *The Chi-squared Distribution*, New York: John Wiley & Sons, Inc., 1969.
- [34] T. Matsuoka and T. J. Ulrych, "Information Theory Measures with Application to Model Identification," *IEEE Trans. Acoust., Speech, Signal Proc.*, Vol. ASSP-34, No. 3, June 1986.
- [35] W. Press, S. Teukolsky, W. Vetterling, and B. Flannery, *Numerical Recipes in C*, 2n ed, New York: Cambridge University Press, 1992.

MULTI-SOURCE DIRECTION FINDING

HRUSHIKESH N. MHASKAR

Professor

Department of Mathematics and Computer Science

California State University
5151 State University Drive
Los Angeles, CA 90032

Final Report for:
Summer Faculty Research Program
Rome laboratories/ERAA

Sponsored by:
Air Force Office of Scientific Research
Bolling Air Force Base, DC

December, 1997

MULTI-SOURCE DIRECTION FINDING

HRUSHIKESH N. MHASKAR

Professor

Department of Mathematics and Computer Science

California State University

5151 State University Drive

Los Angeles, CA 90032

Abstract

We studied the mathematical theory inspired by the problem of multi-source direction finding using a phased array antenna. We studied two approaches for solving the problem: the *orthogonal polynomial* approach and the *wavelet* approach. The first approach provides very accurate predictions, and requires the theoretically minimal number of antenna elements, but is relatively unstable under noise. The wavelet approach requires a large number of antenna elements, which can be simulated by a synthetic aperture radar. This approach also provides accurate predictions, and is extremely stable under noise.

Multi-source direction finding

H. N. Mhaskar

1 Introduction

A phased array antenna consists of several *elements*, each producing an (alternating) electric current when hit by electromagnetic radiation. The problem of direction finding consists of finding the direction of arrival of the radiation from the measurements of the currents. If the radiation is coming from a single source, and the antenna is in a good condition, then there are relatively simple and well understood methods for direction finding. The problem is extremely difficult if there are more than one sources, and one needs to estimate the direction in which each is located relative to the antenna. Because of the great interest in this problem, there are many well known algorithms to solve this problems; a survey can be found, for example, in [3]. However, there is practically no prior work on the mathematical theory governing this problem. The objective of the grant is to begin building such a mathematical theory, which would also lead to better algorithms.

2 Formulation of the problem

Let $k \geq 0$ be an integer. A (narrow-band) radiation incident on the k -th antenna element produces a current of the form $\lambda \exp(i(ku + \phi))$, where $\lambda > 0$ denotes the amplitude of the current, ϕ is the initial phase, and u is related to the direction of arrival θ of this radiation by the formula

$$u = \mu \sin \theta,$$

where μ is a constant depending upon the array parameters and the frequency of the radiation. Therefore, if there are m signals impinging on the antenna at angles $\theta_1, \dots, \theta_m$, with amplitudes $\lambda_1, \dots, \lambda_m$, and initial phases ϕ_1, \dots, ϕ_m , then the resulting current in the k -th element is given by

$$v_k := \sum_{\ell=1}^m \lambda_{\ell} \exp(i(ku_{\ell} + \phi_{\ell})) = \sum_{\ell=1}^m \omega_{\ell} \exp(iku_{\ell}), \quad (2.1)$$

where $u_{\ell} = \mu \sin \theta_{\ell}$, $\ell = 1, \dots, m$. The problem of multi-source direction finding is to find the values of u_{ℓ} , $\ell = 1, \dots, m$, given the vector $\mathbf{v} := (v_0, \dots, v_{N-1})$, where N is the number of antenna elements. We observe that the amplitudes ω_{ℓ} 's can be computed easily once the u_{ℓ} 's are known.

In practice, the observations \mathbf{v} are subject to two kinds of noise. A realistic model is given by

$$\mathbf{x} := (x_0, \dots, x_{N-1}) := s_1(t)\mathbf{v} + \mathbf{s}_2(t), \quad (2.2)$$

where t denotes the time at which the observation was taken, $s_1(t)$ is a complex valued, normally distributed random variable with mean 1, and $\mathbf{s}_2(t) = (s_{2,0}(t), \dots, s_{2,N-1}(t))$ is a vector of complex valued, normally distributed random variables with mean 0. The signal-to-noise ratio in this model at any time depends upon both the magnitudes of the noises s_1 and \mathbf{s}_2 , as well as the amplitudes ω_{ℓ} 's. In our investigations so far, we have adopted the simplifying normalization that $|s_1(t)| = 1$. With this normalization, our model becomes

$$x_k = \sum_{\ell=1}^m \omega_{\ell} \exp(i(ku_{\ell} + \psi_{\ell}(t))) + s_{2,k}(t), \quad (2.3)$$

where ψ_{ℓ} 's are real valued, normally distributed random variables with mean 0, and $\mathbf{s}_2(t)$ is as before.

In summary, the problem is to determine the quantities u_{ℓ} , $\ell = 1, \dots, m$, given $\mathbf{x} := (x_0, \dots, x_{N-1})$. We explored two different approaches towards this problem, which we will call the *orthogonal polynomial approach*, and the *wavelet approach*.

3 The orthogonal polynomial approach

In this section, the Stieltjes integral [7] provides a convenient notation. Let ω be a complex measure that associates the (complex) mass ω_ℓ with each of the points $z_\ell := \exp(iu_\ell)$, $\ell = 1, \dots, m$. For the time being, let us ignore the noises and consider the "pure model" (2.1). Using the notation of the Stieltjes integral, we now see the quantities v_k to be the "moments" of this complex measure:

$$v_k = \int z^k d\omega(z), \quad k = 0, \dots, N-1. \quad (3.1)$$

Let

$$P(z) := \prod_{\ell=1}^m (z - z_\ell) = z^m - \sum_{p=1}^m a_p z^{m-p}. \quad (3.2)$$

Our notation implies that

$$\int f(z) P(z) d\omega(z) = 0 \quad (3.3)$$

for any function $f : \{z_1, \dots, z_\ell\} \rightarrow \mathbb{C}$. Substituting the monomials z^k in place of f , we obtain that

$$\int z^k P(z) d\omega(z) = \int z^{m+k} d\omega(z) - \sum_{p=1}^m a_p \int z^{m-p+k} d\omega(z) = 0; \quad (3.4)$$

i.e.,

$$v_{m+k} = \sum_{p=1}^m a_p v_{m-p+k}, \quad k = 0, \dots, N-m. \quad (3.5)$$

It can be shown that a unique solution of the direction finding problem requires $N \geq 2m-2$. For example, if $m = 3$ and z_ℓ , $\ell = 1, 2, 3$ are *any* distinct points, then *any* observations v_0, v_1, v_2 can be realized with a proper choice of ω_ℓ 's. More generally, let ϕ be any point in the range $0 \leq \phi < \pi/(2m)$, or $\pi - \pi/(2m) < \phi \leq \pi$. It can be shown (cf. [1], Section I.3) that there exist m positive numbers $\lambda_\ell(\phi)$ and distinct points $u_\ell(\phi) \in [0, \pi)$ such that

$$\sum_{\ell=1}^m \lambda_\ell(\phi) R(u_\ell(\phi)) = \int_0^\pi R(\theta) d\theta$$

for any cosine polynomial R of degree at most $2m - 2$. The point ϕ is one of the points $u_\ell(\phi)$. We now consider the points

$$z_\ell(\phi) = \begin{cases} \exp(iu_\ell(\phi)), & \text{if } \ell = 1, \dots, m, \\ \overline{z_{\ell-m}}, & \text{if } \ell = m+1, \dots, 2m. \end{cases}$$

For $\ell = m+1, \dots, 2m$, we write $\lambda_\ell(\phi) = \lambda_{\ell-m}(\phi)$. Finally, we write $\omega_\ell(\phi) = \lambda_\ell(\phi)z_\ell(\phi)^{-2m-2}$. It is not difficult to check that

$$\sum_{\ell=1}^{2m} \omega_\ell(\phi) R(z_\ell(\phi)) = \int_0^\pi R(e^{i\theta}) d\theta$$

for any algebraic polynomial R of degree at most $4m - 4$. In particular, choosing R to be the monomials z^k , $k = 0, \dots, 4m - 4$, we see that any points ϕ in the range $[0, \pi/(2m))$ (or $\pi - \pi/(2m) < \phi \leq \pi$) will give rise to the same observations v_k , $k = 0, \dots, 4m - 4$. In particular, $N = 2(2m) - 3$ is not enough to uniquely determine each of the $2m$ points z_ℓ 's. Actually, the choices $\phi = 0$ and $\phi = \pi$ indicate that one needs $2m$ moments for a unique determination in all cases.

With $N = 2m$, we may apply the equation (3.4) for $k = 0, \dots, m - 1$. Thus, the polynomial P is "orthogonal" to all polynomials of degree at most $m - 1$. This is the reason for calling this approach the orthogonal polynomial approach. We observe, however, that the integral expression in (3.4) is not a proper inner product. Hence, the theory of orthogonal polynomials is not directly applicable here. In particular, the matrix of the system of linear equations (3.5) is not positive definite. Nevertheless, the system is solvable for *almost all* values of ω_ℓ 's.

This suggests the following method for the solution of the direction finding problem. Given the moments v_k , $k = 0, \dots, 2m - 1$, we solve the system of equations (3.5) for the quantities a_p . The roots of the polynomial (3.2) then give us the values of z_ℓ 's, from which the directions can be computed easily. The advantage of this approach is that there are no limitations on how close the angles of arrival can be, except for the limitations of the computer and the algorithms to find the roots of a polynomial. The disadvantage is that the mapping $\mathbf{v} \rightarrow (z_1, \dots, z_m)$ is notoriously unstable, (cf. [2]).

In the presence of noise, one therefore needs a greater value of N than $2m - 1$. The system (3.5) is then overdetermined. We may then take several samples (*looks*) and obtain the v_k 's by averaging. The coefficients a_p can then

be found by a least square fit. Moreover, the singular value decomposition of the matrix provides an educated guess about the number of angles involved. In our experiments, this turned out to be a reasonably stable algorithm, yielding answers with a high accuracy, although the number of looks was very high.

4 The wavelet approach

We use the notation of the Stieltjes integral also to describe this approach. Again ignoring the noise for the time being, let ω now denote the complex measure on $[-\pi, \pi]$ that associates the mass ω_ℓ with the point u_ℓ , $\ell = 1, \dots, m$. The observations v_k can then be written as the trigonometric moments of this complex measure:

$$v_k = \int e^{ik\theta} d\omega(\theta), \quad k = 0, \dots, N-1.$$

The “farfield computation” consists of computing the Fourier sum $\sum_{k=0}^{N-1} v_k e^{-ik\theta}$. This sequence does not converge as $N \rightarrow \infty$. However, for integer $r \geq 1$, the series

$$\sum' \frac{v_k}{k^{r+1}} e^{-ik\theta}$$

(where \sum' means that the term $k = 0$ is excluded) converges uniformly and absolutely to a function with r piecewise continuous derivatives. The points u_ℓ now appear as the points of discontinuity of the derivative of order r of this function.

Wavelets are now extremely popular tools to detect singularities of a function and its derivatives. In our situation, the classical wavelet transforms cannot be computed efficiently. Along with J. Prestin, we developed [4], [5], [6] trigonometric and polynomial frames that are calculated using the moments of the target function, and can detect singularities of all order derivatives. In the present context, our frames take the form

$$\Psi_N(\theta) := \sum_{k=N/2}^{N-1} \frac{v_k}{k^{r+1}} g_{k,N} e^{-ik\theta}, \quad (4.1)$$

where

$$g_{k,N} = g\left(\frac{2\pi k}{N+1}\right)$$

for a suitably chosen function g . We have proved in [6] that the quantity $\Psi_N(\theta)$ is “large near” the points u_ℓ , and “small away” from these. In fact, we have established precise asymptotics for Ψ_N in terms of g .

In our experiments, we used the function g defined on $[0, 2\pi]$ by

$$g(x) := \begin{cases} 0, & \text{if } x \in [0, \pi], \\ (x - \pi)^2(2\pi - x)^2, & \text{if } x \in [\pi, 2\pi], \end{cases}$$

and extended to the whole real line as a 2π -periodic function. The value of r seemed to be irrelevant; we used the value $r = 3$. In contrast to the orthogonal polynomial approach, this method is substantially more stable under noise; we could detect the angles of arrival even with a SNR of $-9.5dB$. It does require a very high value of N . We observe that only $N/2$ values of v_k are used in the calculation of Ψ_N .

To illustrate the relative merits and demerits of the two methods, we considered the incidence angles 10° and 10.01° , with no noise, and amplitudes 1. With 4 elements, the orthogonal polynomial method gave the predictions 10.0093° and 9.9991° . The farfield computation with 256 elements and the frame transform Ψ_{512} (which also uses 256 moments) are shown in Figure 1. (In these figures, the corresponding u -values are on the x axis; the values of interest are 0.5455 and 0.5461 respectively.) It is seen that neither of them is capable of this resolution.

Again, we took the angles 10° and 10.3° , with amplitudes 1. With an additive noise of variance 0.01 and 10,000 looks, the orthogonal polynomial method with 8 elements gave the predictions of 10.1345° and 11.6895° respectively. The farfield computation and the frame transform in Figure 2 were computed with a variance of 3 and just one look. The u -values of interest here are 0.5455 and 0.5617. Clearly, the wavelet approach works much better.

Finally, we took the incidence angles 10° and 15° , with an additive noise of variance 1 and amplitudes 1. With 8 elements and 10,000 looks, the orthogonal polynomial approach gave the predictions of 10.1950° and 15.1915° .

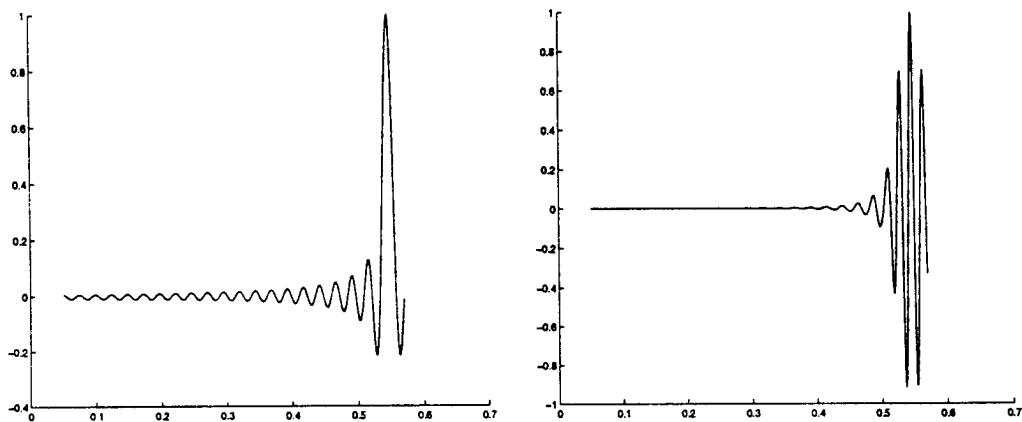


Figure 1. Farfield computation (left) and Ψ_{512} (right) for the u -values 0.5455 and 0.5461, with no noise.

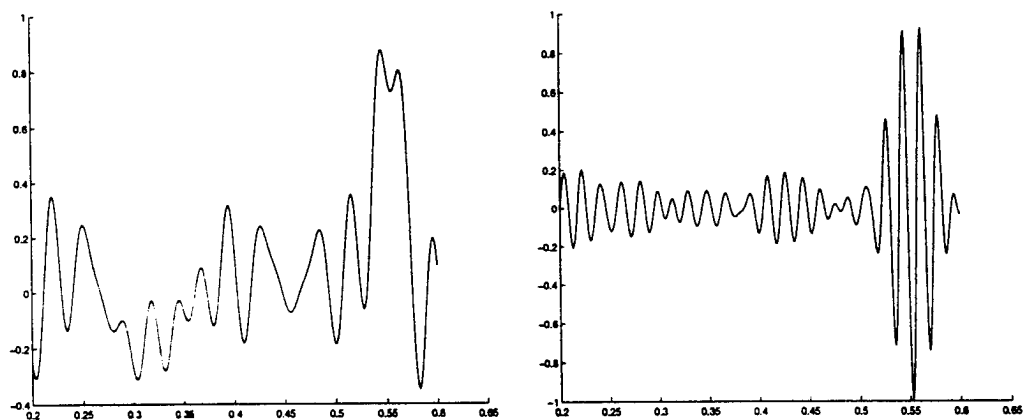


Figure 2. Farfield computation (left) and Ψ_{512} (right) for the u -values 0.5455 and 0.5617, with noise variance = 3.

5 Conclusions

We have studied two approaches for the solution of the multi-source direction finding problem. The orthogonal polynomial approach gives very accurate predictions with a small number of antenna elements, with no limitations on how close the incidence angles can be. However, it is very unstable under additive noise. The predictions are still reliable if the separation is near the higher ranges of super-resolution, but the number of looks required is very large. In contrast, the wavelet approach requires a large number of antenna elements, which can probably be arranged with a synthetic aperture radar, but is very stable under noise, and requires only one look. The frames being trigonometric polynomials, there are theoretical limitations on the resolution that can be achieved, but frames using the same amount of information are capable of higher resolution than the classical farfield computation.

References

- [1] G. FREUD, "Orthogonal polynomials", Pergamon Press, Oxford, 1971.
- [2] W. GAUTSCHI, *Orthogonal polynomials: applications and computation*, Acta Numerica, (1996), 45–119.
- [3] S. U. PILLAI, "Array signal processing", Springer Verlag, New York, 1989.
- [4] H. N. MHASKAR AND J. PRESTIN, *Bounded quasi-interpolatory polynomial operators*, Accepted for publication in J. Approx. Theory.
- [5] H. N. MHASKAR AND J. PRESTIN, *On Marcinkiewicz-Zygmund-Type Inequalities*, To appear in "Approximation theory: in memory of A. K. Varma", (N. K. Govil, R. N. Mohapatra, Z. Nashed, A. Sharma, and J. Szabados Eds.), Marcel Dekker.
- [6] H. N. MHASKAR AND J. PRESTIN, *Polynomial frames for the detection of singularities*, Submitted for publication.
- [7] W. RUDIN, "Principles of mathematical analysis, McGraw Hill, New York, 1976.

AN EVOLUTIONARY SYSTEM FOR MACHINE RECOGNITION
OF SOFTWARE SOURCE CODE

Ronald W. Noel
Assistant Professor
Department of Philosophy, Psychology, and Cognitive Science

Rensselaer Polytechnic Institute
110th 8th Street
Troy, NY 12180-3590

Final Report for:
Summer Research Extension Program
Rome Laboratory

Sponsored by:
Air Force Office of Scientific Research
Bolling Air Force Base, DC

and

Rome Laboratory

January 1999

AN EVOLUTIONARY SYSTEM FOR MACHINE RECOGNITION OF SOFTWARE SOURCE CODE

Ronald W. Noel
Assistant Professor
Department of Philosophy, Psychology, and Cognitive Science
Rensselaer Polytechnic Institute

Abstract

The rapid and efficient development of reliable software, especially large software, is a task for which humans are not particularly well suited. Still, humans possess a broad expertise in programming and software design that has resisted attempts to automate the task. Important to the present study is the human capacity to recognize the intentions of software separate from the code implemented. Recognition of intentions is defined as the ability to categorize software into coherent goal related groups, such as financial, database, or system software. The author in the past created a software evaluation and recognition system from the combination of two recent cognitive science techniques in recognition and categorizing. One technique recognizes objects in a vastly multi-dimensional space using eigenvectors, and the other technique automatically processes text objects using n -grams. The investigation demonstrated how applying these techniques can create a coherent space for recognizing software source code. The present study joined the software recognition system with an evolutionary computational system to create an adaptive and bootstrappable system that shows promise for source code categorization. Such a system could be used in a search or browser system that would immediately benefit to human programmers who review legacy software. Finally, the approach is generalizable to a large domain of evaluation and recognition tasks (such as audio, image, logic and text) because the methods use an informational approach based on atomic (frequency based), not featural (object based), representations.

AN EVOLUTIONARY SYSTEM FOR MACHINE RECOGNITION OF SOFTWARE SOURCE CODE

Ronald W. Noel

Introduction

Software engineering exists in the throes of a cognitive science dilemma. On one hand, the engineering of software is a very human process. It requires broad expertise in programming, software design, and the awareness of the goals and intentions of the software system. For the most part, these human processes are poorly understood by cognitive scientists and have not succumbed to software automatization techniques. The inability to automate such tasks has left software engineering to humans. But, the creation of software, especially large software, is not a task for which humans are well suited. The need to think in exact detail and to consider a large number of simultaneously interacting factors limit the scope, the size, and the reliability of the software that a human or group of humans can produce.

The most reasoned approach to attack this dilemma is the judicious division of software programming tasks between humans and computers. In such an approach, humans would work in broad terms suited to their abilities to manage the goals and intentions of a program. Humans could specify the program in terms of actions and purposes. The specifications of the program could then be turned into code by using formal methods. One would expect that the creation of such code would be of a much higher quality than anything that humans could produce unaided. Indeed, such automated coding from a computer based design management system is the approach that describes much of the work and success of knowledge based system engineering [8, 9].

Further success in raising the quality of software will rely on the computer taking over more and more of the human tasks. The first tasks to consider are those that comprise the better understood cognitive tasks, especially the mundane ones. Automating such tasks would free the humans to work on other tasks. The benefits to software engineering would include greater job satisfaction for the human, a reduction in development time, and an increase in software consistency (i.e., software quality). The level of performance might be lower than that of the human expert for the same task, but the benefits could be obtained as long as the software's performance was near the average for human performance. Additionally, because automated software capabilities are independent of humans, the expertise would not be lost through the attrition of normal work environment. Instead, the software could be developed as a separate component and studied and improved towards the long term goals in computer aided software engineering system.

This paper investigates a system which will recognize the intentions of existing legacy software for use in the development of the functional specifications of a future software. The system will be a pattern recognizer capable of assigning software to coherent goal related groups [10]. Applications for the recognizer might include a filter in a browser, a prefilter that prioritizes software for an automated recovery system, a tool that aides in the segmentation of software into meaningful chunks, and a method for determining the intentions of a software artifact use in automated processes. Consider, when faced with the task of reviewing a vast amount of software, a human would first make a series of quick initial judgments based on holistic pattern recognition to eliminate as many candidates for use as possible, and prioritize any software kept. The remaining software would have to go through some detailed analysis to determine if the software was suitable. Software that passes the quick judgment relies on human holistic pattern recognition ability. Since these are the capabilities that the filter are to emulate a brief discussion of human holistic pattern recognition will be valuable.

Holistic Recognition

A fundamental schism exists between human intellectual abilities and machine processing capacities. Humans are open systems which tend to act in very holistic and non-deterministic ways. On the other hand, machines are closed systems better suited for analysis in a limited domain. Humans lack the tremendous numerical computational speed; yet they can process information holistically in an automatic, rapid, and natural manner. Machines possess tremendous computational capabilities; yet no algorithm exists to perform holistic processes as well as humans do. In general, these differences have led to an antagonistic interface between humans and computers, however, these same differences could lead to a beneficial synergism. Ideally, a good interactive system would integrate the best human qualities with machine computational capabilities enabling it to outperform either of the two cognitive components alone [7].

The theory regarding holistic processing can be separated into two stronger and weaker stances. Under the weaker stance, features (meaningful units separate from context) interact with each other through configural processes to form emergent properties or "second-order relational features". Under the stronger stance, the process is completely holistic in that its representation is non-decomposable, and no explicit description of features exists outside the context. These stances provide two ways to understand systems to support holistic representation: The traditional approach to understanding holism is to seek a system of features and relationship that combine in meaningful and unexpected ways to capture attributes of configurations [11]. A newer approach is to seek a system which endeavors to understanding the relationships between configurations without the decomposition into features [12].

Cognitive research with humans tends to support the use of the newer approach to holism when dealing with expertise. A well-known area in which cognitive researchers study these holistic processes is the recognition of objects and, in particular, faces. Theories regarding the recognition of objects and faces have been distinguished by different perceptual encoding and representational processes. Much of basic object recognition theory has been based on the decomposition of parts and the analysis of edge features [12, 13, 14]. On the other hand, face recognition theory has been based on stronger holistic processes which utilize holistic surface characteristics such as texture, color, and shading [15]. Important to the present paper is that research suggests that the distinctions between object and face recognition begin to fade when one examines expert object recognition[11]. Experts appear to use utilize the stronger holistic processes similar to those found in face recognition.

Machine Categorization of Holistic Representations

Machine categorization of objects is studied under many banners. Each banner represents either a domain of objects to be classified (i.e., images versus text), process (i.e. feature analytic versus template matching), representation (i.e. neural networks versus propositional), or any combination of the three. Considering such broad topics as data modeling, object recognition, and text comprehension one finds that a dominant approach is the decomposing of complex representations into a small number of components or features along with a limited number of possible relationships. The features and their relationships are then captured in a parameterized model that can easily be manipulated or "understood" by algorithms for computer use. For our purposes, a feature-based decomposition method for developing a filter would be to select a set of key words and phases and a set of rules that would categorize based on the presence and absence of the words and phases in a given text.

However, problems occur with such an approach since it is open ended and requires either a human analyst to hand craft the solutions or tailor make automated approaches to pick features and discover rules for relationships. The difficulty of this is that it is usually not clear what, if any, set of features will decompose the representation coherently.

Remember, a representation that cannot be decomposed into a limited set of features and relationships is said to be holistic. Such holistic representations, however, do succumb to atomic or frequency level decomposition, and usually it is the atomic symbols that are manipulated on a machine. The author suggest that software source code which is created by humans is one of these non-decomposable representations when it comes to semantic classification or categorization.

Human-computer interaction commonly uses atomic representation of objects, as shown in the following examples: (1) A word processor stores and manipulates the graphemes or visual symbols of natural language. Through the manipulation of the graphemes, the program enables humans to represent human thought in text. (2) Graphics programs create, manipulate, and store pixel definition of points of light. Through the manipulation of the pixels, the program enables humans to represent images and visual perceptions. (3) Through the manipulation of program primitives, a programmer can tailor a machine actions to achieve some task or intention. In each case, the machine interacts with the human at an atomic level of representation and has no semantic or meaningful representation. The task falls to the human to organize the representations through an abstract understanding of the object and the methods of structuring the atomic elements to represent the objects.

In the programming example, the atomic level of representation in the interaction limits the aid the application can offer the user. The human-computer interface could be improved if computers had a meaningful representation of programs. The machine could be used for

source code data mining or browsing the internet. It could suggest subroutines to accomplish the tasks needed in a program, determine if the program is fulfilling the programmer's intentions, and determine if the programmer intentions are in-line with higher objectives. It could offer help by locating similar programs and determine if the program corresponds with the programmer's intentions.

Eigenfaces in Face Space

Building a machine's capability to understanding objects created from atomic representation requires machine methods that are frequency and not feature based. One such method is contained in the system by Turk and Pentland [2] at the Vision and Modeling Group part of the Media Laboratory at the Massachusetts Institute of technology. This method is used to identify faces from pixel encoded pictures. The method was developed as an extension of earlier work by Sirovich and Kirby [3] to efficiently represent pictures of faces using principle-components analysis. Turk's and Pentland's system creates a face space based on eigenvectors to decompose, store, and recognize face images. The technique is a straight forward use of principle components analysis except for an early step of image compression where the best subspace is built based on coordinates around a small set of exemplar images, termed eigenpictures. The image compression greatly reduces the calculation used in the principle components analysis. Without image compression, the analysis would need to calculate a covariance matrix the size of the number of pixels per picture squared. For instance, the number of pixels (N) in a 256 x 256 image would require calculating a matrix of the size $(65,536)^2$. Determining the eigenvectors and eigenvalues on such a matrix is an intractable task.

The Turk and Penland process for face recognition using low dimension space selecting a training set (M) of pictures and create the vectors $I_1 \dots I_m$ by concating the pixel intensity

values, making sure that the members of the set are normalized. Normalization is required since the analysis will be sensitive to any differences in the data set. For example, if half the pictures have dark background and half have light background then one would expect background light to be a salient dimension in the resulting analysis. Given that background light is extraneous to the task of face recognition, one would want to keep the background light the same for all pictures. A large portion of Turk and Pentland [1, 2] face recognition systems is front-end processing of any picture to remove the background, center the face, scale the face to a set size, etc.

Next, principal components are determined by finding the M eigenvectors and the eigenvalues of the vectors' covariance matrix. The size of the matrix is N^2 . Presently performing principle component analysis of data sets greater than one hundred is difficult. Because most pictures have a data size or pixel number greater than one hundred, one must use a technique for principle component analysis developed by Sirovich and Kirby [3]: if the number of eigenpictures (M) used is less than the number of data points (N), and one subtracts the mean for the data, there are only the $M-1$ degrees of freedom for $M-1$ non-zero eigenvectors. The analysis uses the M eigenpictures to create a M dimensional subspace of the possible images. Encoding into the M dimensional subspace reduces the size of the principle component analysis from N to M , and computational size of the covariance matrix to the size M^2 . Since M (the number of pictures in the training set is usually around 8 to 40) is usually much smaller than N (usually in the tens of thousands), the analysis becomes efficient.

Recognition is accomplished by first transposing a new face into its eigenface components weights. This is analogous to a Fourier transform of the spatial frequencies contributing to the image. Recognition is accomplished by comparing a newly transformed image to the

distance between the image and the exemplars, or other stored patterns. The new pattern is categorized as being similar to the closest stored pattern.

The eigenspace technique has been successful with images but its application to C source code is not immediately apparent. Another approach to categorizing objects by the frequency of its constituent atomic parts is the n -gram techniques. Important for our purposes is that the technique works directly on natural language text. Although the technique uses only the letters of the alphabet and the space symbol it can be directly generalized to C source code when one adds the additional symbols used in the language C. Next, we will examine the N -gram approach.

N -grams

N -grams encoded the atomic symbols of text into n -character sequences. If n equals one then the sequences would equal the single letters in the text, for $n = 2$ the bigrams, 3 the trigrams, etc.. The frequency of the sequences are usually collected by either sliding a "window" of the size of the n -gram across the text one character at a time. For example, the word "example" would generate the sequences "exa", "xam", "amp", "mpl", "ple", "le_", etc. All are sequences in a $n = 3$ n -gram. Using n -grams as a scoring technique allows modeling the statistical nature of a text in a manner that is robust to typographical errors, garbling, etc. Some of the techniques score only the n -gram in a word whereas others score across inter-word boundaries. At small values for n , the n -grams collect mainly information about spelling and word presence, large N 's collect information on common word sequences and thus start representing the grammar of languages. Indeed, the technique using large n -grams can be used to generate random text in the style of the author on which the n -grams are collected. The approach is language independent and can

be used with Japanese and Chinese computer text based on a 16-bit character code to designate symbols.

The approach was developed by Damashek [4] who used a n -gram system to correct spelling and typing mistakes in text. Since then, a complete system for browsing documents has been developed by Pearse and Nicholas [5] with improvements suggested by Crowder and Nicholas [6] to reduce the amount of information required for a usable system. Systems differ in implementation but have the following general features. First, the system collects the n -gram frequencies using a sliding "window" for a text. A n equal to 5 has been found to be adequate for browsing and hypertexting documents [5]. The frequencies are then normalized by dividing the n -gram counts by the total number of n -grams collected. Then, the counts are centered to an average text by subtracting the normalized n -gram frequencies from a corpus of average texts. The obtained vector is the document histogram that may contain thousands of entries. At this juncture one can use the data in several ways.

The Pearse and Nicholas system TELLTAL [5] determines the similarity of one set of text to another based on the normalized document n -gram vectors to calculate the strength of the relationship between the two representations. Also they index text by a query based on a similarity score calculated on whether a document or query contains n -grams. They also offer a method for disambiguating queries by specifying the context for a query. The context emerges from the intersection between two similarity scores: the similarity of the query and the document set, and the similarity of the current document and the document set. The disambiguation of the set requires the ad hoc setting of thresholds for the two similarity judgments.

Eigencodes in Codespace

Both the n -gram and the Eigenspace approaches seek to holistically identify groups of objects. Both approaches seek to capture the frequency of the atomic features in a representation. The difference between the two come from the different techniques used to determine grouping. To date, eigenfaces are formed on dimensions extracted from the data using principle components analysis with grouping determined by vectorial differences in the space, and n -grams determine grouping or proximity primarily through cluster analysis or proximity based on least squared differences of features. Cluster analysis methods are used to identify groups of objects in whatever dimension space the objects are in, whereas factor analytical methods are used primarily to reduce the dimensions of the space the objects are in. Principle components analysis has the advantage of reducing a large number of variables to a smaller number of variables (locations of axis) for further analysis. The reduction, if done appropriately, reduces noise, gives storage economy, and allows the identification of underlying variables or those dimensions on which the C source code varies.

The author has chosen to create a categorization technique for C source code that use the low dimensional eigenspace technique with the frequency counts of the n -gram approach. This should be achievable since both techniques uses counts or intensities of atomic representations in a vectored format. Given that the data are comparable, one may apply the eigenface technique to the n -gram counts of C source code. In other words, to look for eigencodes in a code space and analyze the reasonableness of the space for representing code. The general approach is to make the codespace require the following steps:

- Acquire a small representative set of C source code to form the exemplar set upon which to build an eigenspace for the code.
- Count the n -gram frequencies for each source of code, and place them in a vector.

- Normalize the frequencies by dividing by the total number hits for each vector.
- Center the vector around the average vector.
- Apply large data set conversion if necessary.
- Do a principle components analysis to find the eigenvalues and eigenvectors to describe the lower dimension subspace of the eigenspace.

To test the concept of building a cognitive filter for C source code, the author selected an approach where the representation of the code would be captured in n -grams. The encoding of source code into n -grams allows the code to be expressed in atomic representations. The representation is the set of symbols that are used to form C code and the conditional probabilities of a code being selected given the proceeding $n-1$ symbols. Such a representation is suitable to represent any C source code as well as any general text. The recognition process over the n -grams used the eigenspace low dimensional categorization technique used in the face recognition. This was chosen over the clustering techniques because the eigenspace systems allow a coherent approach to categorization in which the underlining space can be understood in terms of its structure. In particular, one might want to look for a piece of C source code for which one does not have a good example, but can describe by its probable location in the eigenspace dimensions.

A group of eight software programs were selected to test the ability of a filter to derive dimensions for an eigenspace that would result in meaningful categorization of C code. The programs were selected based on the criteria that the programs performed a single function and the function of the code was either system utility, mathematical, statistical, or logical. The users comments and extraneous lines were removed from the code. The programs were then encoded in n -grams with an n of one (i.e., symbol frequency). All possible one-gram symbols were first collected and assigned a position in a vector. Each vector was normalized by calculating the n -gram likelihood for each program by dividing

each n -gram by the total number of n -grams in the program. The average n -gram vector was calculated. A principle components analysis was performed given 7 factors or dimensions for the code space. A characterization of the dimension was achieved by the author by examining the scorings weights for the symbols.

The code space approach to building a low dimensional filter proved to be a promising approach to categorizing software source code. The method produced a coherent, understandable space in which should be advantageous to automated software reclaiming techniques. A surprising finding that the programmer's style is a major dimension detracts some from the coherency of the space. However, that finding is not all bad. The knowledge that a piece of software is programmed in a structured format may well be useful knowledge for any search or post filter analysis of the software. For example, formal methods that decompose legacy software to understand its functions or intentions might work best with structured code. Humans might want to limit searches to structured code so that any software found will be easier to comprehend and to verify as useful.

Evolutionary Systems

The present proposal is to meld the above software recognition system with an evolutionary computational system to create an adaptive and bootstrapable recognition system. The evolutionary system will be used to select the set of n -grams that best differentiate a software space. The creation of such an adaptive cognitive system for the evaluation and recognition of software source code would be an important step towards developing the critical evaluation function for an evolutionary software system that automatically generates software code. Finally, the approach is generalizable to a large domain of evaluation and recognition tasks (such as audio, image, logic and text) because the methods use atomic representations which support holistic processes.

Development of the system relies on an evolutionary computational system of Noel and Acchione-Noel [7] shown to be able evolve holistic objects that use atomic representations. So far the system has only been used to evolve images from humans, but like the eigenface system above the process can be made to accommodate text or code using n -gram representation. Important for the present project is the question of the complexity and the ability of the proposed system to evolve the level of complexity to effectively recognize code. Since questions of the evolutionary process and complexity are crucial to the project a brief description the evolutionary system to be used will follow.

Holistic processes in evolution are the processes of evaluating fitness, selecting mates, and producing offspring that act upon large numbers of simultaneously interacting genes. The strength and order of the interactions preclude the decomposition or pre-definition of gene-to-feature mappings for analysis. The present paper uses a new system that uses humans as holistic judges for evolving images in a representational space that does not pre-define features or gene-to-feature mappings. The system demonstrates human-machine cognition, and gives a direct method for studying and understanding the effects of holistic processes in computation.

Instead of using a feature-based space, the system used a pixel space that effects the resolution of the image space, but forces no dimensions upon the images themselves. The space is based upon atomic or molecular representation, similar to the notions of atomic or molecular decomposition by Fourier Analysis or Wavelets [16]. As championed by the pointillists, small points of just a few colors can be used to create the psychological impression of any form and any color. The use of atomic representation is sub-featural and allows the generation of features along with their configuration. The representation is not constrained at the feature level and encodes a dog, a tree, or a car as easily as a face. For

instance, one could create a space of 25-by-25 pixels with each pixel being any of eight colors. Such a small space has the potential to create an enormous number of images, as many as 2^{1875} . The number of possible images is so large that there exists no real constraints on the variety of forms that may be represented; rather, the model constrains the resolution of the image. The space cannot represent objects that require more than 12.5 lines of resolution in the vertical or horizontal axis, but such a constraint can be reduced by increasing the number of pixels and decreasing the pixels' size.

System implementation required resolving additional issues in the method of reproduction and mutation function. First, usually, simple cross-over points are used as the method of reproduction, but such a linear system is inappropriate for a multi-dimensional space. Instead, we increased the number of cross-over points until the reproductive system considered a cross-over point at every allele. Such a system of uniform crossovers was implemented by randomly selecting between the genes of the two parents with equal probability. Uniform crossovers are thought to be deleterious to evolutionary computation [17], but have been found useful by others [18]. Secondly, if one uses a mutation function that chooses among all possible genes with equal probability for an allele, the mutation function will eventually return the image to a random state. Instead, we limited the mutation function to the gene values of neighboring pixels, causing smaller changes and greater adaptability.

The resulting image evolving system consists of a comma plus system (allows incest) using a population of fifty images in which 10 images are selected by the human for each generation iteration. Note: in a comma plus system the parents are available for selection in the next generation so that each generation after the first is made of parents plus their offspring. The genotype representation is an array of alleles that has the same size as the pixel representation (25x25 pixels). Each allele is a character that corresponds to one of the

possible colors (or genes) for the pixel. Reproduction creates the offspring genotype by randomly and uniformly selecting between the genes of two randomly selected parents at each allele site.

This image evolution provides a new technique for integrating the best qualities of human and machine capabilities to create images. Neither system could produce these images alone. Machines lack the perceptual and memory skills, and humans lack the ability to execute an image holistically. The results show that current theories of evolutionary computation are insufficient to explain the convergence of the images in the absence of a feature-based parameterized space. A new theory of evolutionary computation, the Stochastic Shift Hypothesis, explains how the images converge and provides a basis for modeling atomic-level convergence in a holistic system.

The technique of image elicitation allows humans to use their perceptual and cognitive systems to organize visual noise into the objects of their memories. This process of literally pulling an image out of chaos will affect our understanding of intelligent systems and future investigations across many disciplines. Image elicitation will be useful in studying machine intelligence, as well as in studying top-down processes in interactive intelligent systems. It provides a means for humans to experience how evolutionary computation works by directly immersing themselves in the process. And it provides cognitive researchers with a means of studying human recognition.

The system creates representations with a new level of complexity over previous work in evolutionary computation. The argument for the increases in complexity is based on an increase in the cardinality of the relationships, increases in the number of emergent properties, and an increase in what Löfgren calls interpretation and descriptive processes [19]. In the representation, the potential for complexity is related to the relationships

among the features. In our system, the features are described at an atomic or molecular level (in our case, points of light). The low level of description allows for an extremely large number of relationships to form as compared to methods that search for relationships among high level features, such as a nose or eyes in a face. Our system, which takes the stronger theoretical stance in representing holistic processes, does not describe the high level features; the features themselves must emerge. Having the features emerge results in a greater number of emergent properties, including all of the features and the configuration, rather than the configuration only. Finally, our system has both polygeny and pleiotropy. Because of these relationships between genotype and phenotype, the complexity of both interpretation and description have increased. Löfgren associated those complexities to computational complexity and Kolmogorov complexity, respectively.

Perhaps the most important increase in complexity for evolutionary systems came from our intentions. Current theory is built upon a mind set of parametrized modeling, or reducing complexity of search through decomposition. Our intention was to embrace as much complexity as possible so as to explore a more complex (and natural?) evaluation function. The author believes that advances in the generality and the effectiveness of evolutionary search will come from increases in complexity.

Project

The intent of this project is to create a system that recognizes the intentions of the source code and programmer's comments for the programming language C. The system will create an informational eigenspaces that encodes the code and comments of C software. The system will be self organizing (bootstrapable) and evolutionary because of the extremely large and computational complexity of the problem. Specially, the project seeks to develop the concepts and methods necessary to accomplish the above vision.

In software engineering the intentions for creating a piece of software are separate from understanding the actual algorithm or computation done by the software. For instance, one might have intended to write a program to calculate the interest on a loan given the amount of the loan and the annual percentage rate. However, because of a bug, the program actually calculates something very different. In this case, there is a mismatch between intentions and the actual computation. Also, the intentions may be implemented in a variety of ways with some implementation being more advantageous than others. If one searches for code that performs a computation in a specific manner, one might miss software that does the same computations, but in a different and more advantageous manner.

Consider, a searcher might look for software that calculates the interest on loans, and in doing so finds software that handles numerical problems, such as rounding errors, in actual ways that the searcher never considered. The searcher might then change their intentions for the software to match the better software. Alternatively, the searcher might use the set of software that has the intentions to calculate interest to determine the nature of interest computations. In summary, there are many advantages to the top-down intentions approach, but the approach must be used in conjunction with other systems, human or not, that can determine whether the computations match the intentions.

To garner the advantages of an intentional approach one must find a system capable of recognizing or categorizing intentions. Until recently this was not considered possible since intentions have holism. Holism means something cannot be decomposed into a finite set of features and relations. Specially, intentions are thought to be context dependent and therefore cannot be decomposed, therefore the traditional methods of scientific analysis are not available. However, modern techniques in information compression and evolutionary have developed representations and processes that are capable of holistic recognition.

Approach

The heart on the current approach is the use of principle component analysis to reduce the dimensionality of C source code into a small coherent, meaningful subspace. The subspace, to be meaningful, must position similar C source code in proximity with each other. Code proximity is the result of positioning items in the subspace according to their score on the dimension that make up the subspace. If the dimensions separate the code into meaningful and salient dimensions then the subspace will have a useful coherency that will allow the use of vector distance in logical and fuzzy searches. Given useful coherency then one could use the location in subspace to search, to limit search, and to rank order possible code sources. Noel's (1996) research demonstrated that a principle components analysis based on n -grams in C source code is capable of creating a coherent subspace for C code classification. However, a classification based solely on the primitives of C has limited usefulness to capture the intentions of the software.

A finding of Noel (1996) is that the coherency of C source code can be of two types, structural and semantic. If one created a subspace based on just the code or structural part of programs then coherency would be based on algorithms similarities. For instance, a program that calculates simple interest might be similar to a program that calculates the amount of light reflected from a surface (both are a percent of a number) but both might be different from a program that calculates compound interests (a program that uses either iteration, or growth approximation). On the other hand, if one created a subspace based on the text of variable names, function names, and programmer comments, then programs that use the same words or morphemes (meaningful parts that makeup words) would be close in that space. In such a space, programs that compute simple and compound interest would be next to each other and programs that compute reflected light would be at a distance. The two subspaces could be used simultaneously by using fuzzy logic to conduct

searches or as a way to bridge between the structure and semantics. For example, one might use a text description of what the program is to do to first find matches in the semantic space, then use the matches to bridge into the structure, and possibly back again.

Good N -grams

The n -gram technique is based on super-computing techniques that seek to collect the frequencies of all n -grams in a text up to some window size (usually $n= 5$ to 10). The number of different n -grams grows exponentially with window and sample text size. The possible number is limited only by number of words and the possible order of the words in the sample text. The present technique forestalls the complete use of the n -gram technique in that the number of possible elements that a principle components analysis can use is severely limited. The PC software used in the present project allows up to 100 variables (n -grams) in the principle components analysis. Note that even for main frame computing, principle component analysis software is usually limited to the three to four hundred. This is because an exponential relationship exists between the number of variables and calculation required to perform a principle components analysis. Given that the present technique must use an incomplete set of n -grams, it follows that a method to determine which n -grams to use in the principle components analysis must be established. To this end, one needs to develop the criterion on which to base a selection, and a method to perform the selections.

The first notion upon which a criterion can be developed is the notion of the completeness of a set of n -grams to act as a covering set. The manner in which n -grams are formed leads to each token for a variable (letter or word depending on resolution of the n -grams) being encoded multiple times. In fact, given the sliding windowing used to encode in n -grams each element will be encoded into n separate n -grams. For example, in a letter-level

gram analysis of $n = 3$, one would find the letter “e” in the word “attachment” in the three n -grams; “hme”, “men”, and, “ent”. One could reduce the total number of n -grams to cover a text while maintaining an entry for each variable. However, the reduction requires that one select the n -grams in the set in some way that best maintains the usefulness of the remaining n -grams, or a good estimate of the important sequential structure or syntax amongst the elements. To do this one must define “goodness” of n -grams for our present purposes.

An important issue is the relationship between structure (syntax) and meaning (semantics.) N -grams are a method for capturing the sequence of symbols. The sequence of C source code is directly related to the syntax of algorithm and the comments. This is because intentions have both structure and semantics. Intentions require that persons have both a goal, and a general method to accomplish the goal. For example, If a person has the goal to be mayor of Troy, New York, then intentions come into play only when the person decides to run for the office. Once the person intends to run for major, their behavior becomes organized around the actions necessary to become elected. One could determine or infer their intentions by either listening to their statement about their intentions to run, or watching their actions (running for mayor requires certain sequences of actions.) Since, intentions organize behaviors then one can infer intentions by the structure of a person’s behavior.

The n -grams that the present analysis needs are those n -grams that best flag the differences between groups of source code. What one looks for are n -grams that are common for one group of source code, but are rare for another. The assumption is that the word or word part used to convey intentions should be common in text with the intention, but rare in others. This difference in likelihood of occurrence allows the information metrics, or Shannon’s complexity measure as an evaluation function of gram goodness.

Interim Work

The author used n -grams of length equal to two as the set of variables to measure the frequencies of co-occurrence of C source code which were used to create the eigenspace used in recognition. While useful when dealing with a limited set of vocabulary of C code, the approach failed when scaled to the larger problem of the vocabulary of programmer comments. A large portion of the grant work was focused on the problem of creating a system that could find a limited set of symbols upon which a scaleable system can be created. After many attempts to work directly with n -grams in an evolutionary approach the research turned to reanalyzing the role of n -grams. This rethinking and generalizing of the approach led to a more generalized and less constrained formation of the gram which the author call Rosettas, after the Rosetta Stone used in deciphering Egyptian Hieroglyphs.

While the problem of n -grams that can be used is unlimited, the number of symbol frequencies used in principle components analysis is extremely limited due to the exponential growth in computation. The software used in the research runs on PC's and allows for only one hundred symbols. Principle components packages for larger computer systems are still limited to three hundred to five hundred symbols. Obviously, a traditional n -grams approach could only be complete with n -grams of $n = 1$. In a limited vocabulary such as C source code, one can find an adequate set of n -grams by using only those n -grams of $n = 1$ or 2 that are contained in the sample source code and that further reduce the set by selecting the best particle set of n -grams based on the symbol weightings during multiple runs of the principle components analysis. Trying to scale the system to large n 's the author quickly came to the conclusion that n -grams are far too many and the diagnostic value of any one gram was far too limited to form a viable approach. At this juncture, the use of n -grams was reconsidered.

Semantic N -grams: Rosettas

As stated before n -grams can be considered a type of covering set for text. The set of n -grams, $n = 1$, that comprise the alphabet will cover all words, meaning that for any word, there is a set of n -grams that will cover all letters in the word. This covering is complete, but with limited and hard-to-interpret results. For instance, one that might use the letter “a” more frequently than another text, but for no discernible reason. A recognition system built upon such differences would not be robust when used to categorize new text samples. As the n becomes greater (particularly when $n = 5$ or greater) the interpretability increases, but so does the need for a large number of different n -grams. Basically, for our purposes N -grams are computational too large.

The problem revolves around trying to capture all specific instances that lead to a general instance. Idiosyncratic words, consistent misspellings and unusual prose in diaries were used to connect Kaczynski to the Unibomber’s manuscript. What is useful in the individual or specific case is not useful in general case. For instance, for most cases dealing with meaning one would not care if one spelled “color” or “colour.” The distinction between the author being British or American may be of little use in discovering the intentions of the software but would cause the need for multiple n -grams, and may cause unwanted dimensions in the codespace.

N -grams cover too much. Much of the written English is redundant and of little diagnostic value. The word “the” occurs frequently in our language, but adds nothing to the meaning of text. While the gram “the” may cover a large portion of all text, it is of little use. Further, n -grams cover too little. The gram for “the” might occur frequently, but is of little use in a generalized recognition system. N -grams capture the sequential nature of text, and therefore do not cover components of the meaning that do not occur sequentially.

For instance, consider the verb tense of prose. Past tense might be indicated by the morpheme “ed”, but also by the various irregular verbs like “was”, “had”, “were”, etc. A single n -grams by its sequential nature cannot capture past tense even though past tense n -grams have a high likelihood of co-occurring in the same text. If one loosens the nature of a n -grams to allow the encoding of patterns that co-occur in non-sequential ways, then one could create a structure that could encode multiple sequences that relate to a primitive semantic feature such as verb tense. This possible relaxed structure is the idea that forms the heart of a rosetta. Specially, a rosetta is a sequence of letters (or symbols) that differentiates two sets of the stimuli (arbitrary groups of C source code files) by proportionally covering larger segments of one set of stimuli than the other.

Consider a rosetta that could encode primitive text semantics. The rosetta “was had went ed were” might be related to the tense of a text that is written about the past. The rosetta “good better best ist most” might encode text that is written to praise something. And the rosetta “she her mrs miss ms tress” might encode a text that is written about a female. As can be seen the parts that make up a rosetta would not normally occur in sequence in a text. However, the parts of the rosetta could be spread across a text that contains the semantic component encoded by the rosetta.

The result of moving the encoding of syntax to that of primitive semantic levels should increase the ability of the recognition system to be sensitive to more complex semantic structures. The problem remains as to how to find and evaluate such rosettas.

Methodology

To use an evolutionary approach to compute a rosetta requires defining what a rosetta is in terms useful for evolutionary simulation. Therefore it is necessary to define a genotype,

phenotype, fitness function and the evolutionary systems capable of evolving rosettas. The requirements are that the rosettas be large in terms of length, and capable of containing letters, morphemes, and words that exist across a text that encode a meaningful component of the text's meaning.

The chromosome of the rosetta was selected to have a length of 30. Each gene can be ascribed to any of the 26 capital letters of the alphabet plus a token to represent a space. This required the preprocessing of text to capitalize all lower case letters, and replace numbers and punctuation with a space. The length of 30 was selected based on the following intuitions. First, the length of the rosetta should be long enough to be contain enough different words and word parts so as to be sensitive to the present of it's encode meaning in a text. For instance, a rosetta that tried to encode feminine gender by just "mrs" would not be sensitive to feminine gender in texts that do not communicate martial status. Second, the length should be longer than 26 so that the initial generations could contain all possible letters. And third, a length too great could increase computational time without necessarily increasing the quality.

For each iteration, two groups of text are randomly selected in the following manner. First the number of different texts that makeup the first group is pseudo-randomly drawn by the computer with an equal probability function of 1 though $N - 1$ (N being the total number of texts, which is 8 in this study.) The number of texts in the second group is pseudo-randomly drawn from a equal probability function of 1 through $N -$ (the first group size). The actual texts are then selected randomly to fill the groups. Note: the eight text used are from the previous study (Noel, 1996.) The selections are made pseudo- randomly without replacement. The population size for each generation was selected to be 250 of which one fifth (50) were selected to be parents for the next generation.

The fitness, or ability of the rosetta to be useful is measured in terms of information metrics (i.e. the ability to differentiate groups of code). The evaluation function is the amount of information difference that a rosetta produces between two sets of C source code text. Informational difference is measured as the average difference in information needed, the number of times the rosetta is applied to cover a text group, or to cover two groups of text. Computer software was developed to evolve rosettas on PC computers in the above manner.

Results

The evolutionary time required to evolve a rosetta as described above is about 75 generations. The average wall clock time to evolve one rosette on a thirty-five megahertz 486 IBM PC with unoptimized software was about twelve hours. This speed of computation required fifty days to evolve all 100 rosettas for the first iteration of forming a codespace. The computer time was seen as a potential issue for the systems so efforts to reduce the time were undertaken. At present, the evolution of one rosetta is under thirty minutes. Also, the computation of the rosettas can be done in parallel so the time to evolve all 100 rosettas would be the same.

Figure 1 shows the Eigenvalues of a Principle Components Analysis with Varimax rotation of the 100 rosettas. The first three factors form the best factors upon which to build the codespace. The three factors account for 70 percent of the trace while being both coherent and general dimensions. The factors relate to using structured programming techniques, file and system processing. The other factors simply separated one program from the rest of the examples.

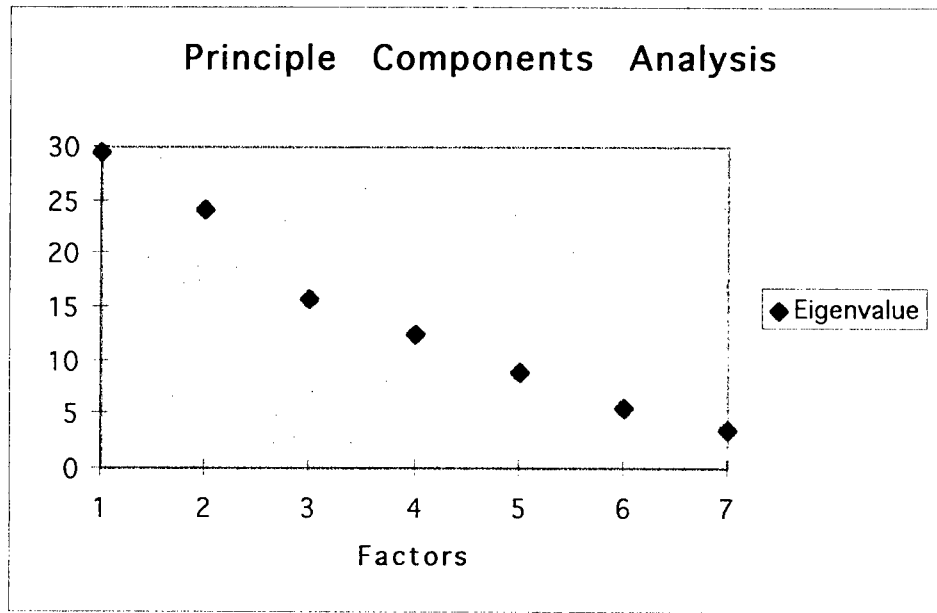


Figure 1. Eigenvalues for vector extraction during principle components analysis

Inspection of the rosettas reveal that the evolutionary approach was able to find patterns that contain letter sequences related to the intentions of the C source code. For instance, examine the rosetta in Figure 2 that loaded highest in determining if the source code involved file processes. The letter sequence “CLO” would be sensitive to the use of the word “close”. The letter sequence word “FL” is used in both “flush” and “flag”. The sequence “NAM” is used in “name” and “filename”. And, the sequence “FILE” stands by itself. Three points can be made: (1) it would be hard to write C source code that intends to do file processes using Standard IO that did not use the underlining words, and so it would be coded by this rosetta, (2) these sequences are unlikely in source code that did not do file processes, and (3) the rosetta appears to be coherent in that it relates to just file source code.

CLOPCFLWTELSUYTFNAMKFILEOARAX

Figure 2. The rosetta that loads highest on the factor encoding file processes.

Conclusions

The progress made in achieving an evolutionary system to recognize intentions is promising, but limited. To summarize, the initial idea of using n -grams to encode C source code and programmer comments and then the creation on a codespace (eigenspace) to recognize intentions was a failure. The failure was due to the high computational demands of large n -grams, and the insufficiency of small n -grams to accomplish the tasks. As a result, an alternative approach was tried. The approach used a new concept called a rosetta. The rosetta is a sequence of letters (or symbols) that seeks to differentiate two sets of the stimuli (arbitrary groups of C source code files) by proportionally covering larger segments of one set of stimuli than the other. The fitness, or ability of the rosetta to do this is measured in terms of the reduction of information (i.e. the increased ability to differentiate groups of code). The usefulness of a rosetta to capture semantics is later determined by it's weighting in forming a factor in an Eigenspace.

The findings are from the development of one set of a hundred rosetta and the formation of an eigenspace from that set. As such, findings are a proof of concept of the ability of an evolutionary system to find rosettas, and the ability of some of those rosettas to encode the stimuli in a meaningful way. What is missing is a demonstration that such a system can act in unsupervised ways and refine its abilities to resolve intentions.

Subsequent activities by the author have moved towards the creation of such a system. These activities included building a software spider to search the Internet to find C source code. A senior Computer Science student, Michael Corbett, was funded for one half year by the author with further funding by Anderson Consulting Company to examine the web spider possibilities. The student developed a shareware spider that he modified to look for C source code on the web. His findings where not promising in that web sites do not normally contain source code, and that seeking such code outside the immediate web site might involve ethical considerations. Additional funding for the project was obtained by including the project in the formation of an applied cognitive science laboratory at Rensselaer Polytechnic Institute called the Minds and Machines laboratory. Continuance of this project is funded by the institute's funding for the laboratory. Also, further programming and work is being through graduate level class projects.

REFERENCES

- [1] M. Turk and A. Pentland, "Face processing: Models for recognition," *Intelligent Robots and Computer Vision VIII*, Proc. SPIE Vol. 1192, (1989), pp. 22-32.
- [2] M. Turk and A. Pentland, "Recognition in face space," *Intelligent Robots and Computer Vision IX: Algorithms and Techniques*, Proc. SPIE Vol. 1381, (1990), pp. 43-54.
- [3] L. Sirovich and M. Kirby, "Low-dimensional procedure for the characterization of human faces," *J. Opt. Soc. Am. A*, Vol. 4, No. 3, Mar. (1987), pp. 519-524.
- [4] M. Damashek, "Gauging similarity with n -grams: Language-independent categorization of text," *Science*, Vol. 267, Feb. (1995), pp. 843-848.
- [5] C. Pearce and C. Nicholas, "TELLTALE: Experiments in a dynamic hypertext environment for degraded and multilingual data," *J. Am. Soc. Inf. Sci.*, Apr. (1996).
- [6] G. Crowder and C. Nicholas, "Using statistical properties of text to create metadata," *Proc. of the First IEEE Metadata Conference*, Apr. (1996).
- [7] R. Noel and S. Acchione-Noel, "Holistic processes in evolution: Evolving non-decomposable images using evolutionary computation," (Unpublished manuscript).

- [8] C. Green, D. Luckham, R. Balzer, T. Cheatham, and C. Rich, "Report on a knowledge-based software assistant," *Final Technical Report RL-TR-83-195*, Rome Laboratory, Aug. (1983)
- [9] M. Gerken, N. Roberts, and D. White, "The knowledge-based software assistant: A formal, object oriented software development environment," *Proc. IEEE Nat. Aero. & Elec. C.*, May (1996), pp. 511-518.
- [10] M. Chase, D. Harris, S. Roberts, and A. Yeh, "Analysis and presentation of recovered software architectures," *KBSE 96*, In Press (1996).
- [11] V. Bruce and G. W. Humphreys: "Recognizing objects and faces." *Visual Cognition 1* (2/3): 1994, pp. 141-180.
- [12] D. Marr and H. K. Nishihara: "Representation and recognition of the spatial organization of three-dimensional shapes." *Proceedings of the Royal Society of London B200*: 1978, pp. 269-294.
- [13] I. Biederman: "Recognition-by-components: A theory of human image understanding." *Psychological Review 94*: 1987, pp. 115-147.
- [14] S. Ullman: "Aligning pictorial descriptions: An approach to object recognition." *Cognition 32*: 1989, pp. 193-254.
- [15] C. J. Price and G. W. Humphreys: "The effects of surface detail on object categorization and naming." *Quarterly Journal of Experimental Psychology 41A*: 1989, pp. 797-828.

- [16] Y. Meyer: Wavelets: Algorithms and applications. (Translated and revised by R. Ryan). Society for Industrial and Applied Mathematics, Philadelphia, PA: 1993, p.4.

- [17] D. B. Fogel: Evolutionary computation: Toward a new philosophy of machine intelligence. IEEE Press, New York: 1995, p.57.

- [18] G. Syswerda: Uniform crossover in genetic algorithms. Proceedings of the Third International Conference on Genetic Algorithms. J. D. Shaffer (Eds.), Morgan Kaufmann Publishers, Los Altos, CA, 1989, pp. 2-9.

- [19] L. Löfgren: Complexity of descriptions of systems: A foundational study. International Journal of General Systems 3: 1974, pp.197-214.

**RAPID PROTOTYPING OF SOFTWARE RADIO SYSTEMS USING
FIELD PROGRAMMABLE GATE ARRAYS**

Glenn E. Prescott

Associate Professor of Electrical Engineering

Department of Electrical Engineering & Computer Science

University of Kansas

1013 Learned Hall

Lawrence, KS 66045

Final Report for

Summer Faculty Research Program

Rome Laboratory

Sponsored by:

Air Force Office of Scientific Research

Bolling AFB, DC 20332

and

Rome Laboratories

Griffiss AFB, Rome, NY 13441

December 1997

RAPID PROTOTYPING OF SOFTWARE RADIO SYSTEMS USING FIELD PROGRAMMABLE GATE ARRAYS

Glenn E. Prescott

Associate Professor of Electrical Engineering

Department of Electrical Engineering & Computer Science

University of Kansas, Lawrence, KS 66045

Abstract

Field programmable gate arrays (FPGA) are powerful new technology which can be used to maximum advantage in military software radio applications. The objective of this research is to examine the potential role of the FPGA in the implementation of high performance military radio algorithms. A radio transceiver implemented using state-of-the-art DSP technology - often referred to as *software radio* - requires real time signal processing at a variety of bandwidths. In order to accommodate the needed bandwidths in a discrete time implementation, it is appropriate to use devices which are well suited to each stage of the system - fast, yet algorithmically simple devices for the wide bandwidth stages and slower, yet more flexible devices for the processing required at low bandwidths. This report briefly discusses the processing requirements of software radio, and assesses the role of the current generation FPGA technology in implementing the algorithms required to make these radio systems function efficiently. A case study is provided of a digital filter design using FPGAs.

RAPID PROTOTYPING OF SOFTWARE RADIO SYSTEMS USING FIELD PROGRAMMABLE GATE ARRAYS

GLENN E. PRESCOTT

1. Introduction

Until recently radio transmitters and receivers were almost exclusively implemented with analog electronic components. However, a new approach is now becoming popular - one that employs digital electronics to implement most of the analog signal processing functions in the radio. This evolution in radio system design is driven by the ever increasing speed and decreasing cost of microprocessors and high performance analog-to-digital (ADC) and digital-to-analog (DAC) converters. It is no longer uncommon to sample a received signal at the intermediate frequency (IF) stage and process the signal with numerical algorithms using a specialized digital signal processing (DSP) hardware. The DSP hardware performs a variety of operations on the signal including down conversion, demodulation, and filtering; all of which are inherently continuous-time (i.e., analog) processes.

1.1 Digital Signal Processing: Capabilities and Requirements

The mathematics of digital signal processing provides the framework for the design of software radio algorithms, while modern high speed digital electronic components make real time implementation of these algorithms possible. However, the hardware currently available to implement DSP algorithms for all stages of the radio system is still limited in speed, accuracy and flexibility. Initially, digital signal processing was used only for baseband waveform processing. As digital electronic devices increased in speed, DSP was soon applied to signal processing functions performed at higher frequencies - e.g., the final IF stage in a radio receiver. Functions such as IF bandpass filtering, automatic gain control (AGC), and coherent modulation and demodulation are typically required at this stage. In the absence of a sufficiently high speed processing capability, innovative techniques such as sub-sampling are used to process bandpass

signals of small to moderate bandwidth. This has allowed the boundary between analog and digital processing to be pushed as far up the signal path towards the antenna as permitted by physical electronic devices. For most types of moderate data rate communications - on the order of 100 kB/s or less - bandwidth is not a serious barrier to DSP techniques. However, military radio systems pose a notable challenge because of the wide bandwidth characteristics of spread spectrum modulation.

1.2 Military Radio Signal Processing Requirements

Military communication systems often require the use of spread spectrum techniques to provide an antijam (AJ) capability; or some measure of covertness through the use of low probability of intercept (LPI) waveforms. The result is that extremely wide bandwidth signals are present at the output stage of the transmitter and the input stages of the receiver. We know from the Nyquist theorem and fundamental bandpass sampling techniques that bandpass signals can be sampled at a rate no less than the bandwidth of the signal; so high frequencies alone do not put a limitation on DSP processor capability. However, wide bandwidth signals are a challenge for any type of digital signal processing hardware, and they are especially troublesome for conventional DSP microprocessors. While conventional DSP microprocessors are optimized for real-time data processing, they are nevertheless implemented using the traditional von Neumann architecture - an inherently serial architecture which uses a single multiplier and executes one instruction at a time. While providing the advantage of flexibility through programmability, this architecture limits the speed with which signal samples can be processed. Even modern DSP microprocessors operating at 40 million instructions per second (MIPS) have a useful bandwidth limit of less than 500 kHz. This is especially troublesome for military communication systems which employ AJ and LPI waveforms having typical bandwidths in excess of 10 MHz.

1.3 Advantages of Specialized Digital Hardware

When digital signal processing at wide bandwidths is required the radio designer turns to specialized hardware which can operate at much higher throughputs than is possible with a DSP

microprocessor. These include application specific standard products (ASSP), application specific integrated circuits (ASIC), and field programmable gate arrays (FPGA).

Application Specific Standard Products (ASSP) such as FIR filters, correlators, and FFT processors, permit certain popular DSP algorithms or functions to be optimized in hardware at the cost of flexibility. Use of ASSPs can significantly increase the device count and often presents special interface problems which can lead to further complications. Furthermore, due to a narrow range of applicability, many ASSPs may not be available in state of the art process technology [1].

When performance is a factor and product volume is high, many designers turn to ASIC technology. ASIC technology offers the ability to design a custom architecture that is optimized for a particular application. For example a conventional DSP microprocessor has only a single multiply-accumulate (MAC) stage (see Section 3), so each filter tap must be executed sequentially. An ASIC implementation of a DSP algorithm, on the other hand, might have multiple parallel multiply-accumulate (MAC) stages. When comparing the performance of the ASIC versus the DSP microprocessor it becomes apparent that the DSP microprocessor offers slow speed but maximum flexibility (due to programmability) while the ASIC provides high speed with minimal flexibility. Between these two extremes lies the field programmable gate array [2].

1.4 Software Radio

The essential concept of software radio is that most of the analog signal processing operations of the radio transmitter and receiver are implemented with digital hardware using DSP techniques. The placement of the receiver analog to digital converter (ADC) and the transmitter digital to analog converter (DAC) as close to the antenna as possible are distinguishing characteristics of the software radio. In the software radio receiver, the approach often used is to digitize an entire band and to perform IF processing, baseband, bitstream and other functions completely in software [5]. This approach requires the use of high speed analog to digital converters and high speed DSP microprocessors. However, the signal processing requirements for military and commercial radio systems employing high data rate signals or spread spectrum modulation easily

exceeds the processing speeds currently available in off-the-shelf DSP microprocessors. In this case, special purpose DSP hardware, application specific devices and field programmable gate arrays can play an important role.

The motivation for implementing radios in software is that a highly flexible and reconfigurable communication system can be implemented for relatively low cost. The ability to adapt the radio to its environment by changing filters, changing modulation schemes, switching channels, using different protocols and dynamically assigning channels and capacity are features which are impractical to deliver with hardware alone. Since the behavior of the software radio can be changed so easily, defining a particular architecture does not limit the radio to one specific function. Instead, multiple radio systems can share a common front-end analog radio tuner while having independent digital processing for each individual radio channel. [5]

2. Discussion of the Problem: Field Programmable Gate Arrays

Modern field programmable gate arrays can implement functions beyond the capabilities of today's DSP microprocessors. In fact, they have the potential to provide performance increases of an order of magnitude or better over traditional DSP microprocessors, but with the same flexibility [3]. These devices can provide the programmability of software, the high speed of hardware and can be reconfigured in-circuit with no physical change to the hardware. In fact, FPGAs are really "soft" hardware, in that they are a good compromise between flexible all-software approaches which unfortunately limit throughput, and custom hardware implementations, which are more expensive and inflexible [4]. FPGAs offer a powerful approach - an architecture tailored to the specific application. Because the logic in an FPGA is flexible and amorphous, a DSP function can be mapped directly to the resources available on the device. Modern FPGAs have sufficient capacity to fit multiple MACs or algorithms into a single device along with the interface circuitry required by the application - a single chip solution.

Programmable hardware has been available for many years - conventional memory devices are the most obvious example. Various PLDs (programmable logic devices) have long

been used in implementing state machines and "glue" logic, among other things. However, the available devices have tended to have restricted architectures and to be rather small [7]. The last decade has seen a significant change with the introduction of a variety of field programmable gate arrays, as well as an evolution of some PLDs into much larger devices with extended architectures. Essentially, the FPGA is a general purpose programmable logic device consisting of a regular array of cells with distributed routing that can be configured with a specific design by the user, without the need to fabricate an application specific device (i.e., an ASIC) [8].

2.1 Programmable Logic Technology

There are a variety of FPGA architectures available depending upon the manufacturer. However, there is one broad distinction that can be made regarding FPGA structure: the architectures are either *course-grained* or *fine-grained* [7]. The earlier devices were simple arrays of logic gates which were programmable in the field in much the same way as a conventional ROM. These devices are considered fine-grained in the sense that there can be a large number of very simple logic operations which can be interconnected. On the other hand, modern FPGAs have a relatively smaller number of more complex logic cells available.

Other than granularity, FPGAs are differentiated by their chip level architecture and their interchip wiring organization. As an example, the Xilinx 3000 family FPGAs consist of an array of cells called CLBs (configurable logic blocks). Each CLB contains two latches and a function generator as illustrated in Figure 1. The internal connections within the cell and the lookup table in the function generators are determined by configuration bits held in an integrated SRAM. This allows an individual cell to implement quite complex combinational and sequential functions. The routing resources allow the cells to be connected as required, at least in principle. In practice, the problem of routing a congested design is the major obstacle in obtaining highest performance.

FPGAs are just beginning to have a significant impact, although their cost is still relatively high (i.e., hundreds of dollars for the largest devices). Two application areas which traditionally have dominated their use are general purpose gate-level logic support (i.e., glue logic) and emulation of new IC designs. However, FPGA manufacturers believe that their

products will change the way in which digital design is approached in a revolution similar to that engendered by the microprocessor [7]. The fact that FPGAs are now being investigated for use in high speed DSP applications is an indication of the broad impact they may have in digital applications of all kinds.

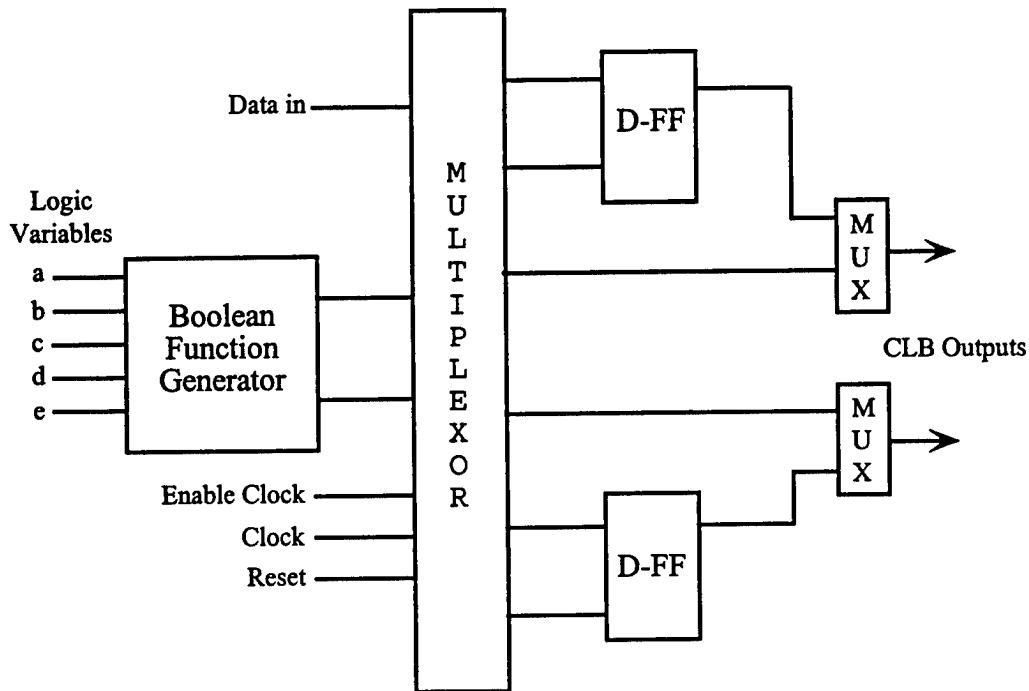


Figure 1 - Configurable Logic Block of the X3000 FPGA

2.2 Practical Consideration in the Use of FPGAs

Because the FPGA is programmable in manner similar to a microprocessor, it is already becoming widely used. However, the configuring of hardware to fit a specific computation is significantly different from the programming of a microprocessor. In particular, the microprocessor has a fixed instruction set, and all solutions are algorithmic in nature. In contrast, an FPGAs internal structure must be customized to implement a particular algorithm. Since digital hardware designs are not software driven, the overhead associated with command interpretation, scheduling and execution is eliminated and there is a substantial gain in speed. Furthermore, a hardware design can take advantage of parallel implementations to eliminate

bottlenecks [2]. It is interesting to note that we may even combine the two approaches and compile a specialized microprocessor into the FPGA with a restricted instruction set chosen to suite any particular application.

It often occurs that a computation is better suited for either dedicated hardware or microprocessor software. This is the situation we are examining in the software radio - when to use FPGAs and when to use DSP microprocessors. Simply stated, an FPGA is appropriate when the design calls for the performance of an ASIC and the flexibility of a microprocessor. An FPGA should not be used if the algorithms to be implemented are complex, or vary significantly in structure or complexity. Determining when to offload DSP algorithms to FPGAs requires an analysis of speed versus problem size. At one end of the scale, problem size gets very large and direct hardware solutions become too difficult and expensive to build [2].

The advantage of FPGAs is that they represent a compact integrated programmable hardware solution which can be user configured for any conceivable logic design. Current designs contain in excess of 40,000 logic gates, all under the control of the designer. On the other hand, FPGAs have some notable disadvantages. First, there internal routing contributes substantial delay between logic elements resulting in a significant limitation in performance, although parallelism and pipelining can still be used. The second disadvantage is that it is not possible to execute a variety of arithmetic operations within the logic resources available. Added to this is that the programming of FPGAs is difficult, especially when implementing DSP functions [9].

2.3 Using FPGAs for DSP Applications

The FPGA has recently generated interest for use in DSP systems because of its potential to implement an infinite variety of custom hardware solutions while still maintaining the flexibility of a conventional programmable device [6]. Although DSP microprocessors have complete algorithm flexibility, their performance is limited because algorithms are implemented by sequential MAC operations, as previously described. Additionally, DSP microprocessors have an overhead for reading in the operands and writing the result through a single data port. Therefore, a DSP microprocessor may require at least four cycles (i.e., read, multiply, add and

write) to perform the simplest of algorithms, resulting in 10 MIPS performance from a 40 MIPS processor [1].

Because DSP algorithms are optimally mapped to the device architecture, FPGA performance can significantly exceed DSP processor performance. For example, a DSP microprocessor can implement an 8-tap FIR filter at 5 Msps. An FPGA can implement the same FIR filter at 100 Msps [1]. FPGAs will never completely replace general purpose DSP processors, however. Current generation programmable logic addresses only the fixed point DSP portion of the market. General purpose DSPs still dominate in floating point performance. Also, general purpose DSP processors utilize familiar software methods, while using programmable logic requires a completely different approach on the part of the DSP designer. Implementing DSP functions in FPGAs provide the following advantages over conventional DSP hardware:

a. *Parallelism* - Using FPGAs can lead to significantly higher performance than a typical DSP processor for some applications.

b. *Efficiency* - An FPGA can be optimized for specific algorithms, thus achieving the performance of hardware with the flexibility of software.

c. *In-circuit Reconfigurability* - Permits the algorithm or function to be changed while operating in-circuit. An additional benefit of FPGAs over ASICs is that they can be reprogrammed on the fly in the system. Consequently, a single FPGA can implement different DSP functions at various times in a system to boost overall performance.

d.. *Adaptability* - A device that can implement large internal RAM blocks can be used to implement real-time adaptive functions at a throughput that cannot be matched by conventional DSP solutions.

2.4 Alternative Arithmetic Options for FPGA

The primary limitation of the FPGA when used in DSP applications is arithmetic - most notably multiplication. When FPGAs are used for DSP applications, the multiplier circuits must be implemented with the available chip resources. However, a hardware multiplier is a reasonably complex circuit, as evidenced by the fact that conventional DSP microprocessors contain only a single hardware multiplier, and it occupies most of the real estate on the chip. A state-of-the-art FPGA can support no more than a handful of multipliers, meaning that brute force multiplication is often avoided in some of the most common operations - e.g., filtering or correlation.

When implementing multipliers in hardware, two basic alternatives are available: The fully parallel array multiplier and the fully bit-serial multiplier. The advantage of the fully parallel array multiplier is that all of the product bits are produced at once which generally results in a faster multiplication rate. The multiplication rate for this adder is simply the delay through the combinational logic. However, parallel multipliers also require a large amount of area to implement. Bit serial multipliers on the other hand generally require only $1/N$ th the area of an equivalent parallel multiplier but take $2N$ bit times to compute the entire product (N is number of bits of multiplier precision) [6]. This concept is illustrated in Figure 2. A new trend is to incorporate a limited number of hardware multipliers within the FPGA. For example, AT&T incorporates a 4×1 multiplier in each programmable function unit of its ORCA FPGA family.

Innovative techniques which avoid conventional multiplication in computing FIR filters and other DSP algorithms have been investigated by a number of researchers. For example use of distributed arithmetic techniques have been reported [12], which makes extensive use of look-up tables, an approach which allows a considerable savings in chip resources.

The FPGA has the ability to implement filtering and transforms using one of several distributed arithmetic techniques, depending on the performance required. These techniques can be used to optimize the implementation of many other types of data processing or MAC-based algorithms. Parallel distributed arithmetic techniques are used to achieve the fastest sample rates, while lower rates can be sustained with a serial or serial sequential distributed arithmetic techniques that uses less resources (fewer arithmetic).

2.4.1 Distributed Arithmetic

Distributed Arithmetic are computational algorithms that perform multiplications with look up tables. This algorithm is generally used to perform important DSP filtering and frequency transforming functions. Since most of the recent architectures of the programmable logic have supported the look-up table methodology distributed arithmetic has become very popular.

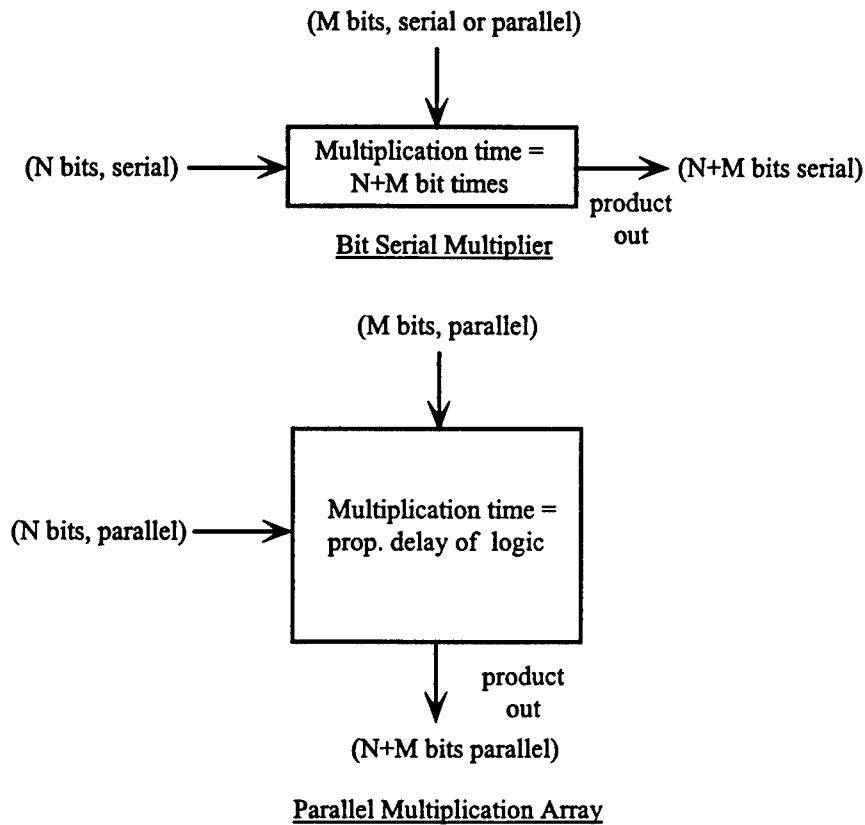


Figure 2 - Block Diagrams of Basic Multiplier Alternatives [6]

Distributed Arithmetic differs from conventional arithmetic only in the order in which it performs operations. Take for example a four-product MAC function that uses a conventional sequential shift and add technique to multiply four pairs of numbers and sum results. The four-multiplication are performed simultaneously and the results are then summed when the products

are complete. This method of implementation requires n -clock cycles for data sample of n -bits. Hence, the processing clock rate is equal to data rate divided by the number of data bits. During each data clock-cycle, the four-multipliers simultaneously create four-product terms, that eventually are summed into the output. The distributed arithmetic differs from this process by adding the partial-products before, rather than after, the bit-weighted accumulation.

By using Distributed Arithmetic, the operations are reordered. The reordering reduces the number of shift-and-add circuits to one, but does not change the number of simple adders. Distributed arithmetic is of two types the serial and parallel distributed arithmetic.

Distributed arithmetic is useful in filtering applications, where the coefficients are constant. Adders and AND gates are made use of to implement multiplication with coefficients. But in distributed arithmetic the AND functions and adders are replaced with look up tables (LUT). If a single bit is made use of to access the LUTs then it is called serial distributed arithmetic, where the incoming sample of the signal stored in a shift register and a bit at a time is shifted out. The other type of distributed arithmetic is Parallel distributed arithmetic (PDA), where the number of bits used to access the LUTs are more than one. The overall performance of PDA is better than SDA as in the former case the number of bits processed during each clock cycle is increased.

2.4.2 Pipelined Architectures

Since the FPGA CLBs contain flip-flops, they are used for storing and delaying the signal. The purpose of pipelining is to increase the speed at which the system operates by decreasing the delay of the critical path of the system. The Figure 3 below illustrates the concept of pipelining.

In this figure we have two clocked circuits which are driven by the same clock. Hence the speed of operation depends on the delay that separates the two clocked circuits. If the delay is more than the clock period, then the frequency of operation is limited by the delay between them, to increase the frequency, register elements can be include in the delay path thereby decreasing the delay in the path and increasing the speed of operation. This is shown in part two of Figure 3. Pipelining is a trade off between speed versus the resource utilization. Using too

many pipelined stages in the system results in a enormous amount of hardware resources and in designs where area and power are of main concern, then pipelining may not be advisable.

2.4.3 Design of Data Paths

The data path is that module of the system that performs the data processing operations. The design of the data path in any system design is very important. The speed of the system increases if there are a number of data paths performing the same function in parallel but this kind of system design requires a lot of hardware resources and hence it is important to decide whether or not the data path will be shared.

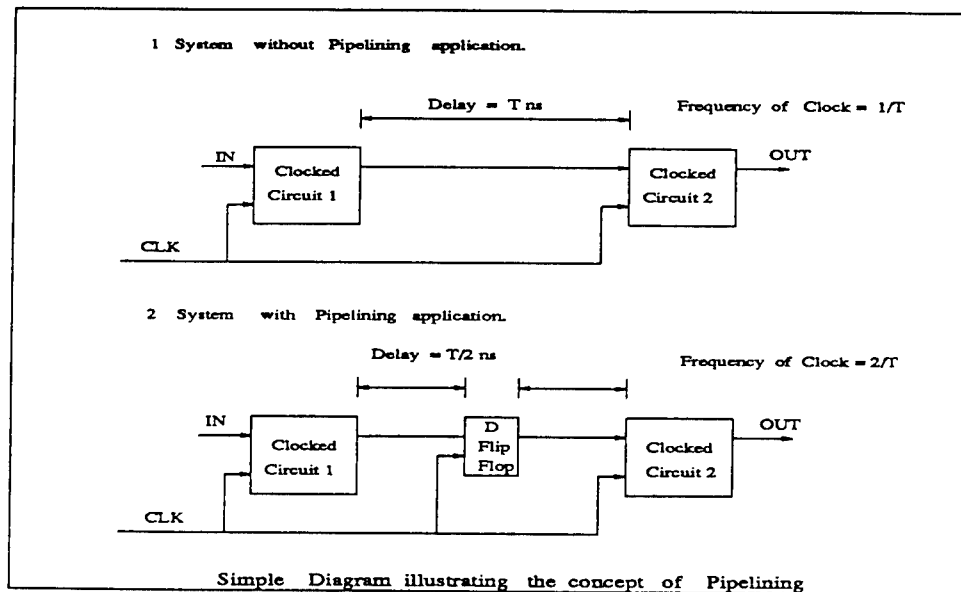


Figure 3 - The Pipelining Concept

A critical factor in designing the data path is the precision requirements of the design. If very high precision is required, especially in DSP applications, then the data path should implement a floating point processing of the signals. However, this would lead to excessive usage of resources since two registers are required to store a floating point number in the data path, hence all the designs make use of fixed point number data paths.

Though fixed point data path design leads to using less hardware resources, the accuracy of the results obtained are less when compared to the floating point designs. This factor is

important for programmable logic as there are limited resources. By making use of fixed point algorithms, a considerable number of resistor elements can be reduced, which in turn can be used to implement parallel data paths and pipelining to improve the speed at which the design operates.

2.4.4 Routing Delays

Routing Delays play a major role in the hardware design since they limit the operating speed of the system. Very complex designs often suffer from routing delays because of the large number of logic circuits implemented in the device, resulting in less space available for achieving optimal routing. This is a concern for DSP systems as many arithmetic operations need to be performed. Arithmetic elements are required which typically occupy a large percentage of the available resources, which affects the routing of the design.

The recent trend in FPGAs is to use common function (or macro) blocks to aid in developing systems with lower routing delays. For example the Xilinx FPGAs supports XBOLX modules such as adders, subtractors, incrementors, decrementors etc. , which are used extensively to implement arithmetic functions on Xilinx FPGAs with very low routing delays. Thus by decreasing the routing delays we can increase the frequency of operation of the system.

2.5 FPGA Applications in Software Radio Systems

The DSP Functions that FPGAs do best are those requiring high sample rates and short word length. They are especially suited for FIR filter designs employing lots of filter taps and fast correlators. The lookup table architecture of FPGAs provides a fast and efficient way to build correlators [3]. More taps can be added to the parallel filter with only a small performance tradeoff with additional parallel silicon resources. In contrast, DSP processors exhibit a linear decrease in performance as the number of taps increases (Table 1). An 8-tap, 8-bit FIR filter implemented on an Altera device needs only 80% more silicon than one 8 x 9 bit fixed multiplier (Table 2)[1].

| TABLE 1 - FULLY PARALLEL 8-BIT FIR FILTER (FLEX 8000A FPGA) | | |
|--|---------------------------|--|
| <i># of Taps</i> | <i>Performance (MSPS)</i> | <i>Equivalent MIPS (DSP Processor)</i> |
| 8 | 104 | 832 |
| 16 | 101 | 1,616 |
| 24 | 103 | 2,472 |
| 32 | 105 | 3,360 |

| TABLE 2 - SILICON RESOURCE COMPARISON | | |
|--|------------------------------------|-------------------------------|
| <i>Function</i> | <i>Inputs & Outputs</i> | <i>Flex 8000A Logic Cells</i> |
| FIR Filter | 8-bit data, coeff 17-bit output | 296 |
| Fixed Point Multiplier | 8 x 9 bit data 17 bit output | 164 |

Table 3 shows the performance of multipliers implemented on the Xilinx 4000 family. Note that parallel multipliers require a larger proportion of the device, while bit serial implementations are slower. The first number in the Multiplier Speed column for the bit-serial multipliers is the clock speed, while the second number is the multiplier speed.

| TABLE 3 - XILINX 4000 SERIES FPGA MULTIPLIERS | | | |
|--|--------|-----------|--------------|
| Type of Multiplier | # CLBs | % of FPGA | Mult. Speed |
| 8 bit unsigned (parallel) | 64 | 16% | 8.54 MHz |
| 16 bit unsigned (parallel) | 242 | 60% | 3.8 MHz |
| 8 bits unsigned (bit-serial) | 17 | 4% | 73.1/4.6 MHz |
| 16 bit unsigned (bit-serial) | 33 | 8% | 62/1.9 MHz |

FPGAs can efficiently implement IIR filters. For example, a lookup table based vector multiplier can be used to create a complete second order section of an all pole analog filter. The

vector multiplier requires the same resources and operates at the same speed as a fixed point multiplier. A Butterworth filter can run at a rate of 25 Msps and require only 139 logic cells [1].

Altera has developed high speed FIR filter megafunctions that are optimized for their own FPGA structure. These filters can be implemented in parallel or serial form allowing a tradeoff between silicon resources and performance. Parallel filters can perform at rates up to 100 Msps enabling digital processing of RF-IF data. Serial filters require less logic and still perform at 5 to 6 Msps. In a Spread Spectrum RF modem application, an Altera FPGA can implement the receiver's correlation filter function at a chip rate over 60 MHz. A DSP processor can perform the remaining tasks, such as quadrature phase shift key (QPSK) demodulation. The resulting DSP application can deliver six times the data rate as the DSP processor alone.

A typical DSP algorithm contains many feed-back loops or parallel structures. The software code for a DSP algorithm of this type is not efficiently implemented in general purpose DSP. Typically, about 10-30 percent of the DSP code utilizes 60-80 percent of the processors power. Analyzing the DSP algorithm and breaking out any parallel structures or repetitive loops into multiple data paths, one can enhance the overall performance of the algorithm. To increase the speed and resource utilization, the multiple parallel data structures can be processed either through parallel DSP devices or in a single FPGA-based DSP hardware accelerator with or without the assistance of a DSP device.

The use of FPGAs are well suited for many DSP algorithms and functional routines. The FPGA can be programmed to perform any number of parallel paths. Hence by implementing the algorithm in parallel paths the speed at which the algorithm can work on a piece of hardware increases, thus one very important use of FPGAs in the field of DSPs is improving the speed of operation. The operational data paths can consist of any combination of simple and complex functions, such as Adders, Barrel Shifters, Counters, Multiply and Accumulation, Comparators and Correlates just to mention a few. The FPGA can also be partially or completely reconfigured in the system for a modified or completely different algorithm.

The primary concept is to unload the compute-intensive functions requiring multiple DSP clock cycles into the FPGA and allow the DSP processor to concentrate on optimized single-clock functions. The combined functionality of the FPGA and general-purpose DSP can

support several magnitudes higher data throughput than two or more parallel DSP devices. The FPGA/DSP implementation is more flexible and proves to be more cost effective than multiple DSPs or an ASIC.

3. Rapid Prototyping Concepts

Designing with FPGAs requires computer assistance at almost every stage of the design including detailed specification, simulation, placement and routing. The use of schematic capture based CAD tools is a common approach to the design of custom logic devices using FPGAs. This process is often combined with logic level simulation to verify a specific design. One method of increasing the range of architectural solutions that a designer may explore in a reasonable time is to specify the DSP system with a hardware description language (HDL) [10]. This steps the design process up one level and allows a generic functional description of the target system which can be further simulated or implemented directly onto an FPGA after the HDL code is converted using the FPGA manufacturers software.

In DSP applications, arithmetic circuitry for operations such as addition, subtraction and multiplication are commonly required. These arithmetic circuits can be designed and implemented by employing user-generated or manufacturer-provided sub-circuits, which can be reused. However, as these designs can only be simulated at the logic gate level, it is difficult to verify the functional performance of the algorithms being implemented. It is particularly difficult to determine the potential undesirable side effects of finite precision arithmetic, as this may require that large data sets be simulated and translated from numerical values to logic levels and vice versa [10]. However, new software tools are being developed which raise the design process to yet another level, allowing the designer to begin at the system level.

Simulation tools such as Cadence's Signal Processing Worksystem (SPW) now have features which allow the engineer to design hardware logic systems and DSP fixed point systems using the traditional block diagram functional description of the circuit. This design is then immediately converted into a hardware description language. Other SPW tools allow the design to be simulated via the HDL description of the system and then linked into a manufacturers

software tools which support specific devices. Most manufacturers, in the interest of making their product more attractive to their customers, have developed a set of stock logic elements which can be reused within their device to assist the engineer in quickly achieving any design.

Once suitable design tools and automatic methods are perfected, designers and programmers will be able to create custom hardware circuitry and pipelines to suit the problem at hand - the term 'soft hardware' suggests that hardware will become as readily created and malleable as software. In a practical sense this will mean that the turn-around time for custom hardware will be just as short as software development is today.

3.1 Design Flow

Figure 3 below gives the flow of rapid prototyping. The flow of the design is from the functional description of the system to hardware implementation. The functional description of the system is done in either SPW(Signal Processing WorkSystem) or in VHDL. The functional description is usually is at a high level i.e., basically proper functioning of the algorithm is given importance. A reconfigurable flat-form is made use of to aid the flow of rapid prototyping. To integrate the above methodology commercial CAD tools are used.

The system to be implemented is functionally described using the Hardware Design System (HDS) of SPW. The algorithm is designed using the blocks available in HDS. The functional description of the system is done using the fixed point blocks to limit the use of hardware resources which are limited in reconfigurable hardware. The algorithm developed is simulated in SPW environment before the implementation is done to make sure that the functional description of the system is correct. To implement the design on hardware the HDS section of SPW provides a link which generates VHDL code for the system designed in HDS. The generated VHDL code is used for synthesis to implement the system on the targeted device.

3.1.1 Functional Description Using SPW and VHDL

The other way to functionally describe the whole system is in VHDL. The advantage of using hand-coded VHDL, rather than the VHDL code generated by the schematic tool is, for very large and complex designs, the schematic capture of the system becomes difficult and

impractical. Also, the hand-coded VHDL is very flexible in the sense that, if certain enable signals are required for flip-flop, counters etc, then the blocks provided by the HDS need to be modified and then used in the system design. However, in the case of hand-coded VHDL we can describe the above components easily. Also the generated code from a schematic tool tends to require more hardware than hand-coded VHDL.

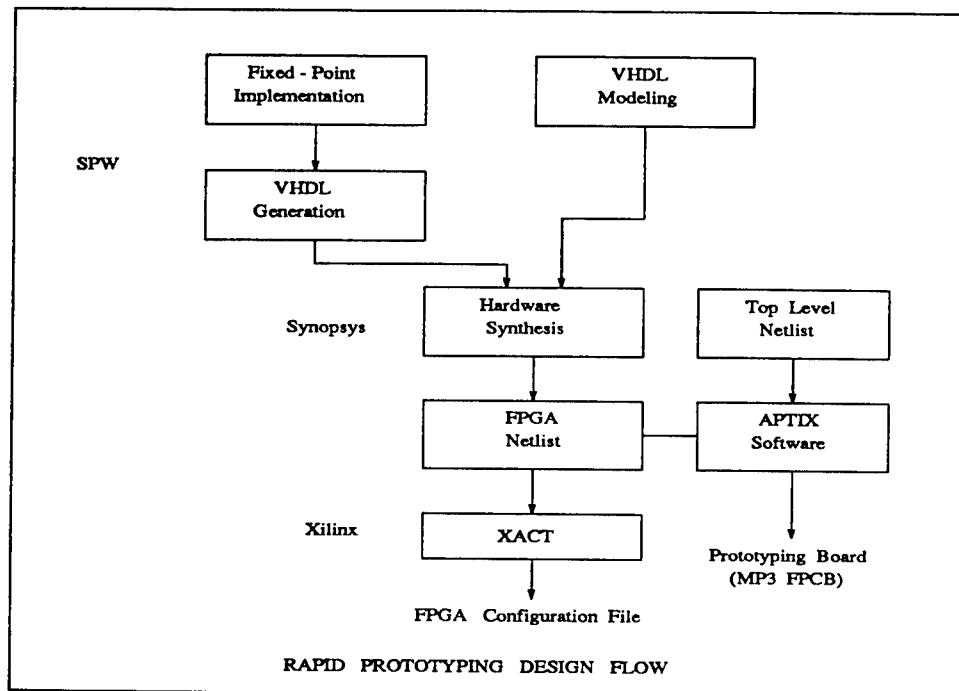


Figure 4 - Block Diagram of Rapid Prototyping Flow

3.1.2 Simulation and Synthesis

The next step in the design flow is to simulate the design using standard tools, that verify the whether or not the design is functionally right. If the simulation results are satisfactory then the design is synthesized using standard tools that target the design to FPGAs. The systems designed using SPW , can be simulated in the SPW environment. But the hand-coded VHDL , need to be compiled and simulated using standard simulation tools of Mentor, Synopsys, etc. The figure 3.2 below shows the flow followed during the simulation of the system.

Synthesis is the procedure that makes possible the implementation of the system on the

targeted hardware. It is also one of the key factors which aids the rapid prototyping. Synthesis tools help in translating the high level design into gate and register level which the routing software understands. The synthesis tools generate netlist files that are used by the routing software to generate files that are used to develop the hardware physically. For example the FPGA Synopsys compiler generates a top level netlist for the design which is used by the Xilinx software, which partitions, places and routes the design. Figure 6 below gives the flow of synthesis procedure followed for synthesizing a given system.

High level system design is gaining popularity as it allows the designers to describe systems at a high level using schematic capture, VHDL and Verilog design. The high-level design methodology reduces library and technology dependence, enabling re-targeting to other libraries, such as an FPGA library, with greater ease.

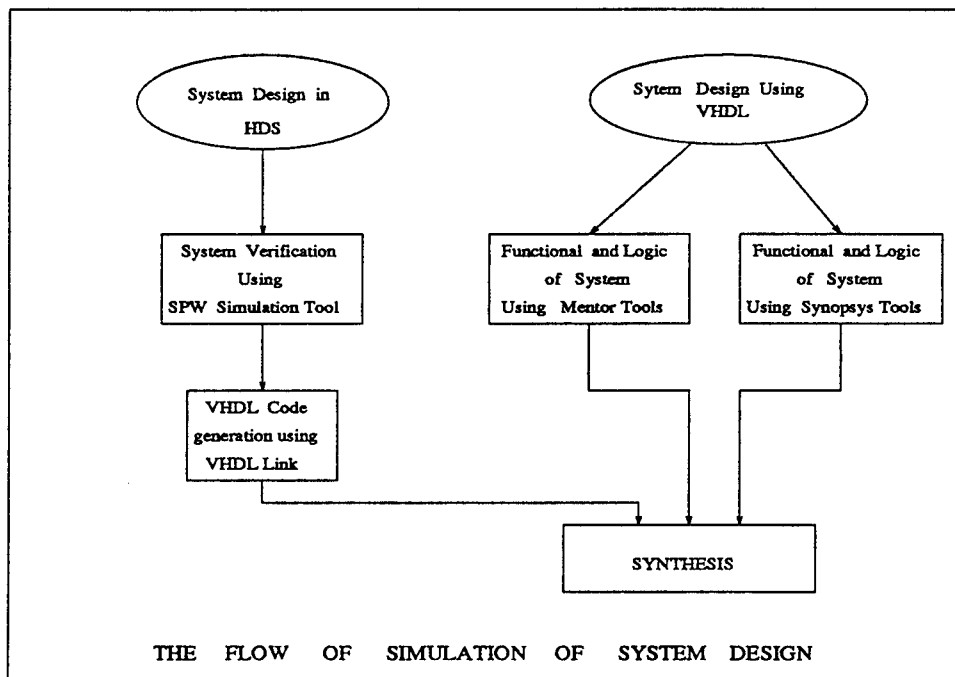


Figure 5 - Block Diagram of Simulation flow

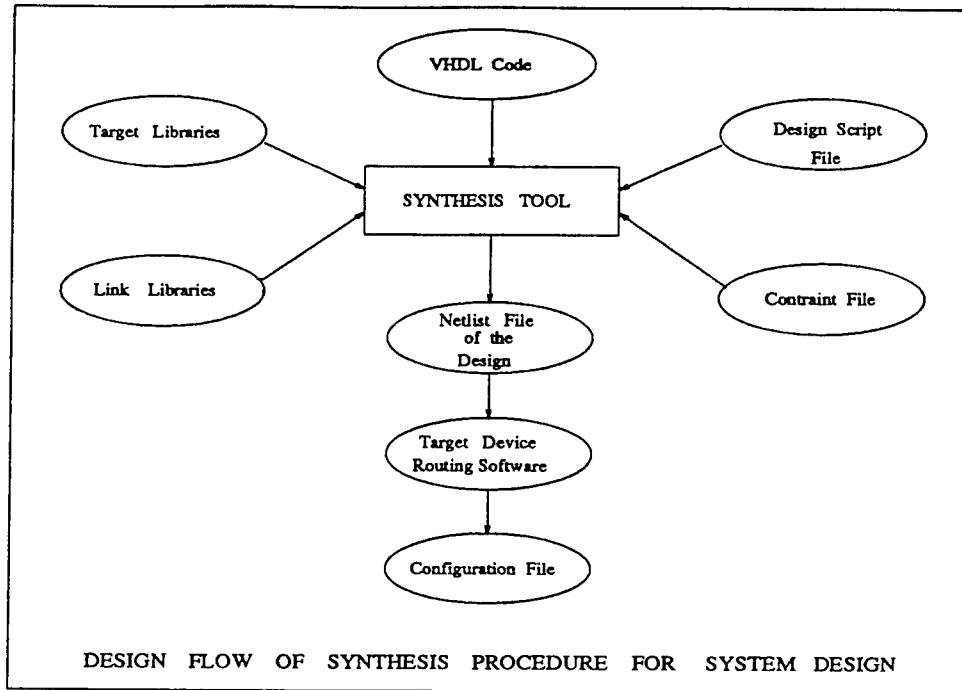


Figure 6 - Block Diagram of Synthesis Procedure

3.1.3 Implementation

The implementation occurs at two levels: first is at chip level where the entire system is partitioned into smaller submodules and these modules are implemented on the FPGAs. The second level of implementation is at the system level, this is where all submodules are integrated and the entire system is tested.

Chip level implementation which is accomplished on FPGAs is an important part of the rapid prototyping flow. Synthesis tools provide the designers with netlist files of the submodules to be implemented. During the synthesis, constraints are provided to meet the specifications. The reconfigurable nature of FPGAs aid the rapid prototyping methodology. The design of the modules are mapped on to the FPGAs using the concept of partition, placement and routing. Every FPGA has its own placement and routing software to map the design. The software partitions the design and places the logic into the configurable logic blocks and finally does the routing of the entire design. The software generates a bit file to configure the FPGA device.

System level design implementation is done using the rapid prototyping board, known as the field programmable circuit board (FPCB). The FPIC (field programmable interconnect component), are programmable interconnect components which form the core of the programmable circuit board. For example, the Aptix MP3 reconfigurable board has three programmable interconnect components used for routing purposes. The MP3 board also supports diagnostic programmable interconnect components which aid in viewing signals on the diagnostic instruments. The FPIC is configured through a Host Interface Module (HIM) , which transfers data from a workstation to program the FPIC. A Stand-alone Program Module (SPM) can be utilized to perform the same function without a workstation.

The FPCB provides fully automated downloading of configuration data to both FPGAs and FPIC devices. As the board supports the combination of FPGAs with standard components(memory, DSP and microcontrollers) makes the MP3 uniquely suited for DSP system prototyping.

3.1.4 Verification

The rapid prototyping environment helps in debugging and verifying very complex systems. As stated earlier, the FPIC devices used on the FPCB are of two types: one is used for routing purposes and is designated as an FPIC(R) device . The other type of device is the diagnostic device, which is used for probing, debugging and verifying signals of the design and designated as FPIC(D). These FPIC(D) devices can be connected to the logic analyzer with help of diagnostic pads. The software for the board is called AXESS, and it programs both the diagnostic FPIC devices and logic analyzer.

This setup provides a very powerful debugging capability, since each signal that appears in the system level netlist can be routed through one or more FPIC(D) devices and viewed on the logic analyzer. The signals to be viewed are selected with the help of diagnostic device interface provided by the software. The software automatically programs the diagnostic FPIC device to display the selected signals on the logic analyzer. At the same time, the diagnostic interface facility configures the logic analyzer. The diagnostic interface provided by the reconfigurable board software does the channel assignment and labeling of the waveform displays.

3.2 Prototyping

In traditional prototyping approaches, the design is mapped to a technology that allows speeds such that all interfaces to targeted applications can operate in real time. But rapid prototyping , provides flexibility for system emulation technology to explore architectural and implementation alternatives available for achieving the desired system function.

Prototyping was commonly done using custom printed circuit boards and wire wrap technologies until the design complexity became too large to make these approaches feasible. The new technologies such as FPIC, FPCB, and FPGA have created a new path that enables mapping of complex logic into programmable hardware which can meet the real-time operating frequencies of DSP applications. The main aim of rapid prototyping is to design, implement and verify systems quickly, hence aiding in bringing products faster to the market when compared to traditional prototyping methods.

4. Methodology: 12 Tap FIR Filter using LUT Techniques

Finite Impulse Response (FIR) filters play an important role in the design of practical discrete-time systems. At the heart of a FIR filter lies the multiplication function, which introduces the coefficients of the filter in the design. Each filter tap has its own multiplier, which gives the product of the input data with the coefficient. When implementing a FIR filter on an FPGA, the multiplication function imposes a bottleneck on the speed performance and area requirements of the design, therefore the designer should focus on enhancing the performance of these multipliers and hence of the whole design. Various multiplication techniques include Shift-and-Add, Adder Tree, Logical Tree, multiplication by a power-of-two, and constant coefficient multiplier using Look-Up-Tables (LUT). The last is the key to high performance in FIR filters with fixed coefficients. This report describes the design of a Low Pass 12-tap FIR filter using constant coefficient multipliers implemented on a Xilinx 4013 FPGA using the Aptix MP3 prototyping board.

4.1 Background

FIR filters are very useful in DSP applications because they are inherently stable, can be employed in a non-recursive structure, and exhibit linear phase characteristics in the passband.

The general Direct Form structure of an FIR filter is shown in Figure 7 below.

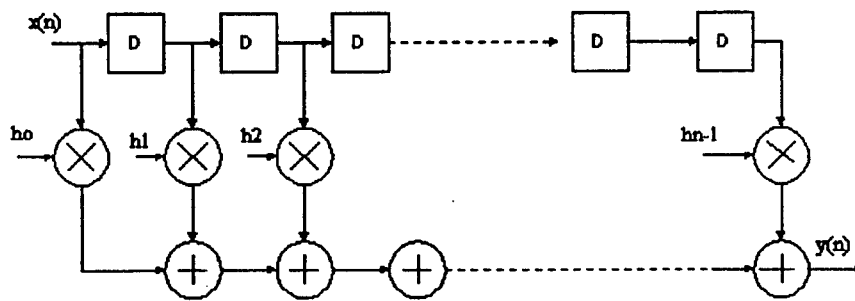


Figure 7 - Direct Form structure of an FIR filter

The algorithmic form of the Linear Constant Coefficient Difference Equation (LCCDE) which describes the system is given by:

$$y(n) = h_0 x(n) + h_1 x(n-1) + h_2 x(n-2) + \dots + h_{n-1} x(n-m)$$

For linear phase response of the filter, the impulse response must satisfy the symmetry condition:

$$h[M-n] = h[n] \text{ for } n=0,1,2,\dots,M$$

The general Direct Form structure shown in Figure 7 exhibits excessive redundant hardware and poor timing characteristics when implemented on hardware. An alternative inverted structure implementation is shown in Figure 8 below.

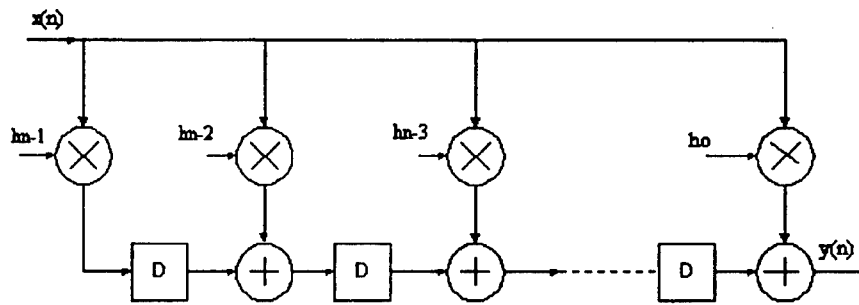


Figure 8: Inverse FIR structure

As shown in the inverse structure, the data samples are applied to all the tap multipliers the same time, therefore processing of the data samples is done in parallel and hence the overall timing performance is enhanced. Also, by exploiting the symmetric nature of the coefficients, we can reduce the number of multipliers needed by half. Moreover, if we use Look-Up-Tables instead of regular multipliers, the time delay incorporated in the multiplication function is dramatically reduced. Each Look-Up-Table in each Tap contains all the possible products obtained when we multiply the specific tap coefficient with the incoming data. Therefore, the data bits are just applied on the address input of the LUT (which is basically a ROM) and the corresponding "data X coefficient" product is obtained automatically on the output of the LUT.

4.2 Design with SPW

The Signal Processing Worksystem (SPW) tool provides the means for designing the system schematically and for verifying and simulating the design. SPW is a powerful block oriented software tool suitable for designing any kind of DSP systems. The Filter Design System (FDS), which is a part of SPW, was used to obtain the filter's coefficients. First, the filter's frequency characteristics (Low Pass, cutoff frequency etc) were given as input to FDS, which in turn calculates the coefficients and the number of taps needed to meet the desired specifications. Then, the Block Design Editor (BDE), which is another subsystem tool of SPW, was used to design the filter schematically using standard DSP blocks like adders, multipliers, delay elements etc.

All these blocks are located in the Hardware Design System (HDS) library of SPW, which allows the use of fixed-point arithmetic in the design. The advantage of using fixed-point arithmetic is that we can accurately model the real behavior of the digital system because we don't need to deal with loss of precision when using floating point arithmetic in a bit-limited digital system. After the system is designed schematically, the Signal Calculator System of SPW is invoked to simulate the operation of the design. The Signal Calculator is capable also of generating fixed-point signals which can be applied to the design and verify its real performance.

4.3 Filter Design

The FIR filter described in this report has the following characteristics:

Type: Low Pass FIR

Tap length: 12

Cutoff frequency: $f_c = 0.1F_s$ (F_s =sampling frequency)

Stopband edge: $0.13F_s$

Stopband Attenuation : 30 dB

Filter Method: Equiripple/Low Pass

Input Data width: 8 bits

Output Data width: 12 bits

Coefficients: 8 bits

Using the Filter Design System (FDS) which is a part of the SPW design tool, the coefficients of the filter were obtained (in Double precision format)

$$b_0 = b_{11} = 0.040473$$

$$b_1 = b_{10} = 0.075372$$

$$b_2 = b_9 = 0.11826$$

$$b_3 = b_8 = 0.16903$$

$$b_4 = b_7 = 0.20653$$

$$b_5 = b_6 = 0.22994$$

Figure 9 below illustrates the block diagram of the design. As shown, the design is implemented as a parallel inverse structure and only 6 Look-Up-Table (LUT) blocks are used because the 12 coefficients are symmetric.

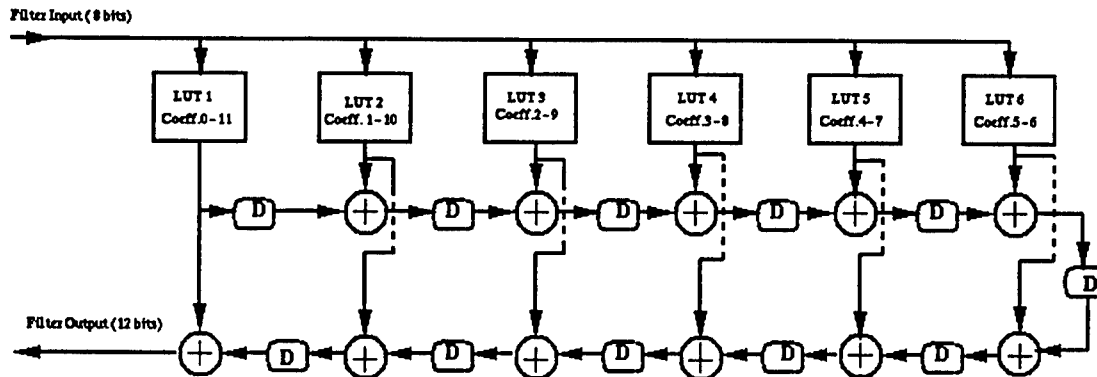


Figure 9 - Block Diagram of the Design

As discussed before, the LUT blocks are used instead of regular multipliers. The internal structure of each LUT block is shown in Figure 10 below. The incoming 8-bit data is split into two segments of 4-bits each. Each 4-bit segment is used to address a ROM Look-Up-Table. So, each LUT block shown in Figure 8 it actually contains two ROM LUTs. We could have had only one ROM LUT and apply all the eight bits of the data on it, but this would be space consuming on the FPGA because it would need a ROM with $2^8 = 256$ memory locations. By splitting the data in two, each ROM has now $2^4 = 16$ memory locations.

The upper (Most Significant) LUT contains all the partial products of the 8-bit coefficient times the most significant 4-bits of the data (i.e 16 partial products). Similarly, the lower (Least Significant) LUT contains all the possible partial products of the 8-bit coefficient times the least significant 4-bits of the data. Therefore, at the output of each LUT we have a 12 bit partial product (4 bits data + 8 bits coefficient). Each 12 bit partial product is appropriately zero padded and then Summed to produce the final product at the output of the Tap.

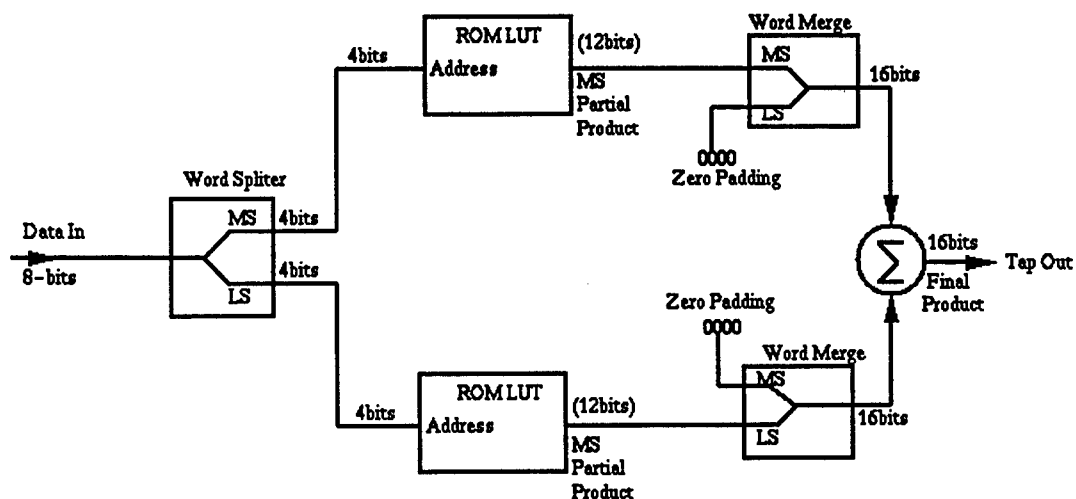


Figure 10 - Internal Structure of LUT

4.4 Results

4.4.1 Filter Performance

The design was implemented on a Xilinx 4013 IO FPGA on the Aptix MP3 prototyping board. Using the Synopsys and Xilinx tools, the FPGA area used and the max time delay of the filter were obtained as shown on Table 4 below. Table 5 is a comparison of the filter described in this report with two other filters designed before.

| TABLE 4 - AREA AND TIMING PERFORMANCE RESULTS | |
|---|------------|
| Total number of CLBs used | 302 |
| % space of 4013 FPGA used | 52% |
| Max. data arrival time (=max delay) | 51.65 nsec |
| Max. filter clock speed | 20 MHz |

| TABLE 5 - COMPARISON OF AREA & TIMING PERFORMANCE OF 3 FILTER DESIGNS | | | |
|--|------------------|-----------------|------------------|
| Filter Type | Total # of CLB's | Max. Data Delay | Max.filter clock |
| FIR 8-tap with regular multipliers | 140 (24%) | 87.8 nsec | 11.4 MHz |
| FIR 12-tap with regular multipliers | 329(57%) | 130.36 nsec | 7.5 MHz |
| FIR12-tap with LUT constant coefficient multipliers | 302(52%) | 51.65 nsec | 20 MHz |

As shown from Table 5 above, the LUT based FIR filter has much better speed performance compared with the two other filters which use regular multipliers in the design. Even the 8-tap filter, which uses much less space than the 12-tap LUT based filter, has bigger time delay in the Data path compared with the 12-tap LUT based design.

5. Results of the Implementation

Figure 11 below shows the theoretical frequency response of the filter obtained using Double Precision arithmetic for the coefficients. On the other hand, Figure 6 illustrates the experimental frequency response with fixed point coefficients. From Figure 5 we observe that there is a steep roll-off at the cut off frequency ($0.1F_s$). At this cut off point we can see that the magnitude of the response falls by -3dB from the maximum. This graph was obtained using Double Precision arithmetic for the coefficients which means that it gives us the desired frequency response with the characteristics given from the Filter Design System tool. In Figure 12 however, we observe that the real frequency response of the filter differs from the desired theoretical response of Figure 10. The -3dB point on the experimental response occurs at a normalized frequency of $0.08F_s$.

As shown on Figure 12, the -3dB point occurs at around 0.08 F_s and not at 0.1 F_s which is the design specification. This happens mainly because of the quantization error introduced on the coefficients. According to our design, the coefficients of the filter are represented by an 8-bit fixed point binary number. Therefore, when the Double Precision coefficient number is represented by the 8-bit fixed point number, there is loss of precision because of quantization. This error can change the filter's characteristics, and especially the frequency response and cut off point, because the original value of the coefficients changes after quantization. For example, coefficient b_0 has a value of $b_0=0.040473$. When this number is represented with 8-bit fixed point (two's complement) arithmetic, it becomes 0.0390625 which is the closest approximation to the original number. Other than that, the frequency response of the filter shows good rejection characteristics and meets the -30dB rejection in the stopband as specified in the design.

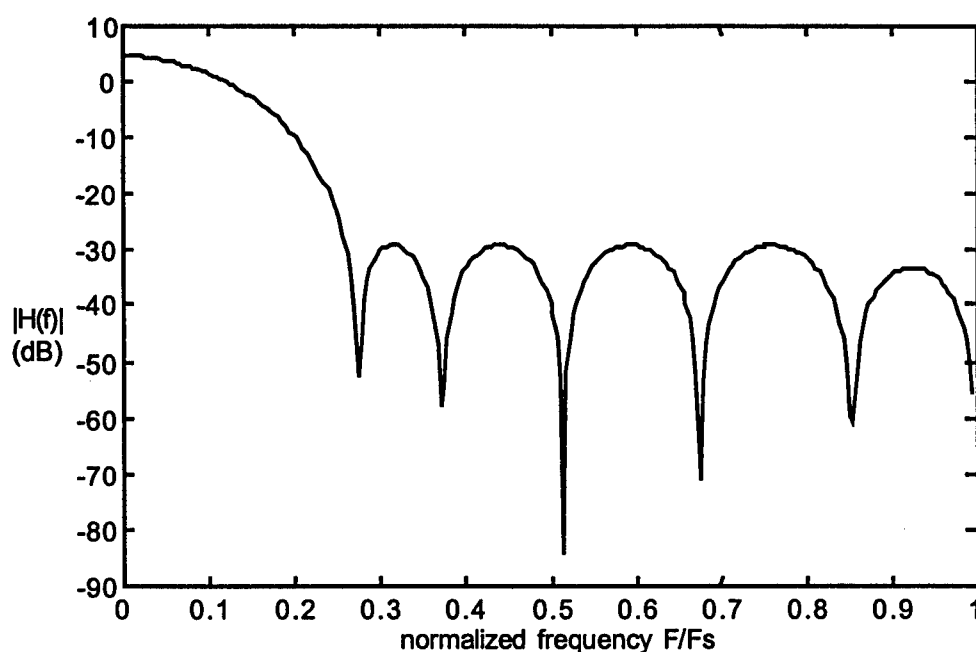


Figure 11 - Theoretical Frequency Response

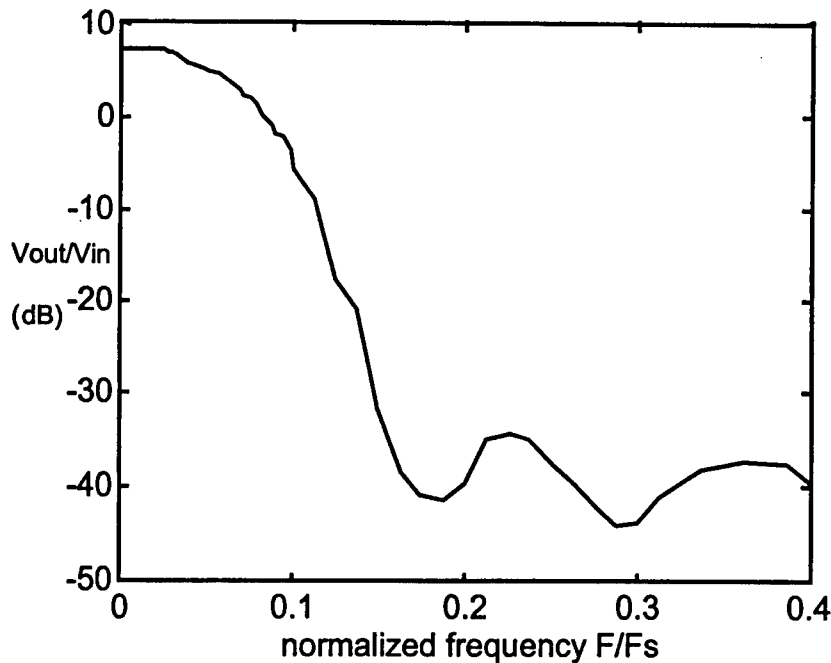


Figure 12 - Experimental Frequency Response (fixed-point coefficients)

6. Conclusions and Recommendations

This report has briefly examined the emerging field of FPGAs and their application to traditional DSP problems. We compared the computational power of the FPGA to that of the DSP processor, and we found that FPGAs are more suited to simple algorithms executed at high speed, while DSP microprocessors are better suited for slower, more complex tasks. DSP microprocessors are limited by their serial-based architecture, while FPGAs can implement any architecture which may be optimum for a particular algorithm. On the other hand, we recognized the shortcomings of the FPGA - these devices are essentially still in their infancy, from a technological standpoint, and there are problems to be solved by the manufacturers before FPGAs seriously challenge DSP microprocessors in the DSP market.

Designing with FPGAs is still a difficult task. The software necessary for efficient FPGA design is extremely expensive and rare. Often, a considerable amount of gate-level modification is necessary on the part of the engineer to realize the performance that the FPGA is capable of

delivering. The current cost of FPGAs is an indication of their novelty in the marketplace. Prices will need to drop by several orders of magnitude before FPGAs will be a common commodity in the DSP designers bag of tricks.

We also presented the design of a Low Pass 12-tap FIR filter using Constant Coefficient Multipliers implemented in Look-Up-Tables. Compared with filter designs which use regular multipliers, this filter exhibited much better speed performance and area occupation on the FPGA. A maximum delay of 51.65nsec was obtained which allows the filter to operate on clock speed of 20 MHz. The frequency response of the filter shows good rejection characteristics (-30 dB in the stop band), but due to quantization error introduced on the coefficients, the cut off frequency is shifted from $0.1F_s$ to $0.08F_s$.

This design occupies 52% of a Xilinx 4013 FPGA, which means that there is still enough space on the FPGA to expand the design with more taps. The benefit of this will be better frequency response characteristics with a trade off on delay increase since the data will have to travel in longer paths. Investigation of this expansion is planned for the future. Also, future work includes investigation of some other design techniques for delay reduction (pipelining for example). Other types of filters are also under investigation (High Pass, Band Pass etc).

7. References

- [1] Digital Signal Processing in FLEX Devices, ALTERA Product Information Bulletin 23, See <http://www.altera.com>.
- [2] Leo Petropoulos, "Replace Digital Signal Processors with HCPLDs," *Electronic Design*, September 5, 1995, pp. 99 - 104.
- [3] Steven Knapp, "Using Programmable Logic to Accelerate DSP Functions", Xilinx Application Note, 1995.

- [4] Eric. L. Upton and Thomas J. Kolze, "Reconfigurable Modems Serve as Multi-Application Communications Node Integrators," *1993 Conference of the American Institute of Aeronautics and Astronautics*, pp. 1 - 3.
- [5] Rupert Baines, "The DSP Bottleneck," *IEEE Communications Magazine*, May 1995, pp. 46 - 54.
- [6] Russell Petersen and Brad Hutchings, "An Assessment of the Suitability of FPGA Based Systems for Use in Digital Signal Processing," *Field Programmable Logic and Applications: Proceedings of the 5th International Workshop*, FPL-95, Oxford, United Kingdom, August/September 1995, pp. 293 - 302.
- [7] A. Lawrence, A. Kay, W. Luk, T. Nomura, "Using Reconfigurable Hardware to Speed up Product Development and Performance," *Field Programmable Logic and Applications: Proceedings of the 5th International Workshop*, FPL-95, Oxford, United Kingdom, August/September 1995, pp. 111 - 117.
- [8] S. Kotta and S. Simanapalli, "Rapid Prototyping of a Digital Signal Processor," *Field Programmable Logic and Applications: Proceedings of the 5th International Workshop*, FPL-95, Oxford, United Kingdom, August/September 1995, pp. 844 - 847.
- [9] Paul Dunn, "A Configurable Logic Processor for Machine Vision," *Field Programmable Logic and Applications Proceedings of the 5th International Workshop*, FPL-95, Oxford, United Kingdom, August/September 1995, pp. 68 - 77.
- [10] L.E. Turner and P.J.W Graumann, "Rapid Hardware Prototyping of Digital Signal Processing Systems using Field Programmable Gate Arrays," *Field Programmable Logic and Applications Proceedings of the 5th International Workshop*, FPL-95, Oxford, United Kingdom, August/September 1995, pp. 129-138.

[11] Joe Mitola, "The Software Radio Architecture," IEEE Communications Magazine, May 1995, pp. 26 - 38.

[12] Bernie New, "A Distributed Arithmetic Approach to Designing Scalable DSP Chips," EDN, August 17, 1995, pp. 107 - 114.

**WAVELET TRANSFORM TECHNIQUES FOR ISOLATION, DETECTION &
CLASSIFICATION OF CONCEALED OBJECTS IN IMAGES**

**Raghuveer M. Rao
Associate Professor
Department of Electrical Engineering**

**Rochester Institute of Technology
79 Lomb Memorial Dr.
Rochester NY 14623-5603**

**Final Report for:
Summer Faculty Research Program
Rome Laboratory**

**Sponsored by:
Air Force Office of Scientific Research
Bolling Air Force Base
Washington DC**

and

Rochester Institute of Technology

December 1997

WAVELET TRANSFORM TECHNIQUES FOR ISOLATION, DETECTION & CLASSIFICATION OF CONCEALED OBJECTS IN IMAGES

Raghuveer M. Rao
Associate Professor
Department of Electrical Engineering
Rochester Institute of Technology

ABSTRACT

The wavelet transform decomposes a signal or image into components on the basis of position as well as scale. Self-similarity across scales is an essential attribute of fractal properties of images. Also, a scale analysis of projections of objects provides feature vectors that are invariant to scaling and rotation. Both fractal analysis and scale analysis of projections have been found to possess good discriminatory properties in the isolation, detection and classification of concealed objects in images. The wavelet transform provides a natural framework for the implementation of these analyses. In fact, the only known wavelet matching techniques for optimum, matched filtered pattern recognition are for bandlimited wavelets. Accordingly, this research effort investigated the development of techniques for fast implementation of bandlimited wavelets and the development of a systems framework for scale analysis of discrete signals and images. The investigation has led to the development of perfect reconstruction circular convolution filter bank structures, construction of continuous-dilation, linear, scale-invariant, discrete-time self-similar signals and systems, and formulation of the scaling mixture problem for multichannel angle and doppler estimation for wideband signals.

WAVELET TRANSFORM TECHNIQUES FOR ISOLATION, DETECTION & CLASSIFICATION OF CONCEALED OBJECTS IN IMAGES

Raghuveer M. Rao

1 Introduction

The wavelet transform decomposes a signal or image into components on the basis of position as well as scale. Self-similarity across scales is an essential attribute of fractal properties of images. Also, a scale analysis of projections of objects provides feature vectors that are invariant to scaling and rotation. Both fractal analysis and scale analysis of projections have been found to possess good discriminatory properties in the isolation, detection and classification of concealed objects in images. The wavelet transform provides a natural framework for the implementation of these analyses. In fact, the only known wavelet matching techniques for optimum, matched filtered pattern recognition are for bandlimited wavelets [4]. Accordingly, this research effort investigated the development of techniques for fast implementation of bandlimited wavelets and the development of a systems framework for scale analysis of discrete signals and images. The investigation has led to the development of perfect reconstruction circular convolution filter bank structures, construction of continuous-dilation, linear, scale-invariant, discrete-time self-similar signals and systems, and formulation of the scaling mixture problem for multichannel angle and doppler estimation for wideband signals and the results are presented under the appropriate headings.

2 Perfect Reconstruction Circular Convolution (PRCC) Filter Banks

Our research into fast implementation structures for bandlimited wavelets has led to what we call *perfect reconstruction circular convolution* (PRCC) filter banks [1, 2, 3, 8]. These filter banks are designed entirely in the discrete Fourier transform (DFT) domain and satisfy perfect reconstruction properties in the discrete frequency domain. The formulation of these filter banks is made on the assumption of a finite length input sequence of length N . In addition, all the operations associated with the filter bank

implementation, namely, filtering, downsampling and upsampling can be implemented entirely in the discrete frequency domain. To see how this can be done, we first interpret the processes of downsampling and upsampling in terms of the DFT. Note that, during this discussion we use the notations $X(k)$ and $X(e^{j2\pi k/N})$ interchangeably to denote the DFT of an N -length sequence $x(n)$.

2.1 Downsampling

Consider Figure 1 which shows an M -fold downsampler. The input to an M fold downsampler is a sequence $x(n)$ of N samples. The output of the downsampler, denoted by $y(n)$, is given by [7]

$$y(n) = x(Mn) \quad (1)$$

where the sequence $y(n)$ has N/M samples. It is assumed that N is an integer multiple of M .

To obtain an expression for the DFT of the output sequence $y(n)$, we first define an N -length intermediate sequence

$$x_1(n) = \begin{cases} x(n) & n=\text{multiple of } M \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

Clearly

$$y(n) = x(Mn) = x_1(Mn) \quad (3)$$

Now the DFT of $y(n)$ is

$$Y(e^{j\pi k/(N/M)}) = \sum_{n=0}^{N/M-1} y(n)e^{-j\pi k/(N/M)}, \quad k = 0, 1, \dots, N/M - 1 \quad (4)$$

Substituting (3) into (4) gives

$$Y(e^{j\pi k/(N/M)}) = \sum_{n=0}^{N/M-1} x_1(Mn)e^{-j\pi k/(N/M)}, \quad k = 0, 1, \dots, N/M - 1 \quad (5)$$

Since $x_1(n)$ is zero at all n except when n is a multiple of M , (3) is equivalent to

$$Y(e^{j\pi k/(N/M)}) = \sum_{m=0}^{N-1} x_1(m)e^{-j\pi k/N}, \quad k = 0, 1, \dots, N/M - 1 \quad (6)$$

Now, we can represent $x_1(n)$ in terms of $x(n)$, we define a comb sequence

$$c_M(n) = \frac{1}{M} \sum_{l=0}^{M-1} e^{j\pi ln/M} \quad (7)$$

Observe that $c_M(n)$ the Inverse DFT of a sequence of M 1's and hence is periodic with a period of M . Thus,

$$c_M(n) = \begin{cases} 1 & n=\text{multiple of } M \\ 0 & \text{otherwise.} \end{cases} \quad (8)$$

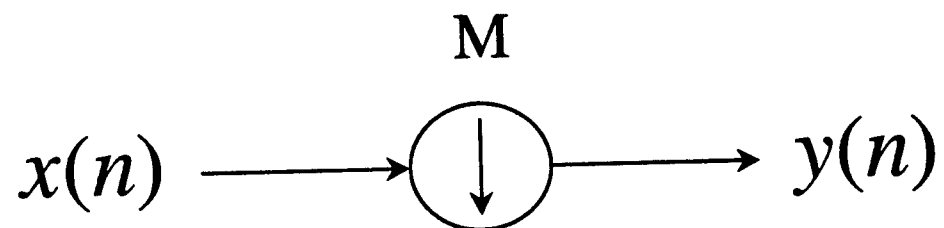


Figure 1: An M -fold downsampler.

Using (7) and (8) we can now write

$$x_1(n) = c_M(n)x(n) \quad (9)$$

or

$$x_1(n) = x(n) \frac{1}{M} \sum_{l=0}^{M-1} e^{j\pi ln/M} \quad (10)$$

Substituting (10) into (5) gives

$$Y(e^{j2\pi k/(N/M)}) = \frac{1}{M} \sum_{n=0}^{N-1} x(n) \sum_{l=0}^{M-1} e^{j2\pi nl/M} e^{j2\pi kn/N}, \quad k = 0, 1, \dots, N/M - 1 \quad (11)$$

which can be simplified to

$$Y(e^{j2\pi k/(N/M)}) = \frac{1}{M} \sum_{l=0}^{M-1} X(e^{j2\pi(k - \frac{N}{M}l)/N}) \quad (12)$$

$k = 0, 1, \dots, N/M - 1$. The derivation given above is very similar to that presented in [7] for general sequences. Note that $Y(k)$ is N/M periodic. In other words, the IDFT of the first N/M samples of $Y(k)$ give the downsampled version of the input sequence. From (12) we see that the DFT of the downsampled sequence $Y(k)$ is a sum of M coefficients of the DFT of the input to the downsampler, $X(k)$, spaced N/M coefficients apart. For example, when $M = 2$, the steps involved are the following:

- Take the DFT of $x(n)$
- Add the DFT of $x(n)$ and its $N/2$ rotated version. This makes use of the N periodicity of $X(k)$.
- Divide the resulting sequence by 2.
- Take the IDFT of the first $N/2$ samples.

This creates the $N/2$ point downsampled sequence, $y(n)$.

2.2 Upsampling

The block diagram of an upsampler is shown in Figure 2. An L -fold upsampler, where L is an integer, takes an input sequence $x(n)$ of N samples and produces an output

$$y(n) = \begin{cases} x(n/L) & n = \text{multiple of } L \\ 0 & \text{otherwise.} \end{cases} \quad (13)$$

The output sequence $y(n)$ has NL samples. The DFT of $y(n)$ is given by

$$Y(e^{j2\pi k/(NL)}) = \sum_{n=0}^{NL-1} y(n) e^{-j2\pi kn/(NL)} \quad (14)$$

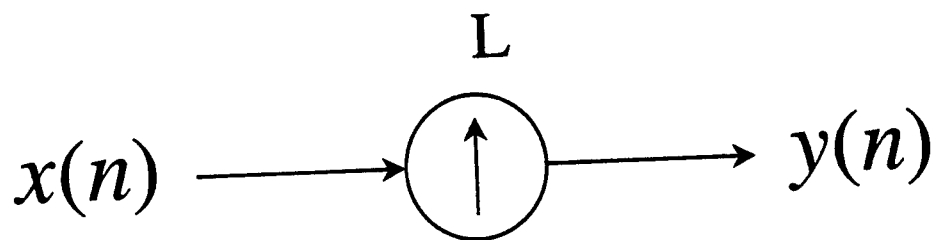


Figure 2: An L -fold upsampler

$k = 0, 1, \dots, NL - 1$. Noting that only each Ln^{th} , $n = 0, 1, \dots, N - 1$ sample of the upsampled sequence are non-zero, we can write (14) as

$$Y(e^{j2\pi k/(NL)}) = \sum_{n=0}^{N-1} y(Ln)e^{-j2\pi kn/N} \quad (15)$$

$k = 0, 1, \dots, NL - 1$. This can be simplified to

$$Y(e^{j2\pi k/(NL)}) = X(e^{j2\pi k/N}), \quad k = 0, 1, \dots, NL - 1 \quad (16)$$

In words, the NL length DFT of the upsampled sequence $Y(e^{j2\pi k/(NL)})$ is nothing but a concatenation of L DFTs of $x(n)$. The derivation of (16) is again along lines very similar to the derivation given in [7] for an infinite sequence.

2.3 Procedure for Implementation

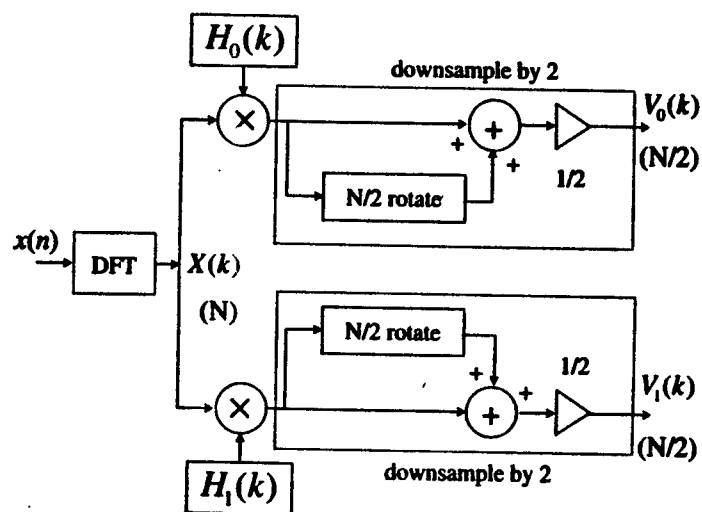
PRCC filter banks are filter banks designed and implemented completely in the *discrete frequency domain* and satisfy the conditions for perfect reconstruction over a discrete set of frequencies. The operations of downsampling, upsampling and filtering are carried out entirely in the discrete frequency domain. The basic procedure for their implementation can be understood by referring to the Figure 3. The analysis side of the system is shown in Figure 3(a). Here, we first take the DFT of the N length input signal $x(n)$. Next, we multiply this DFT, $X(k)$, pointwise with a sequence $H_0(k)$, which is the DFT of an N length sequence $h_0(n)$. This amounts to circularly convolving sequences, $h_0(n)$ and $x(n)$. The resultant sequence is then downsampled by two as explained in section 2.1. The procedure is repeated for the lower branch with $H_1(k)$, the DFT of $h_1(n)$. In this manner, we decompose the input sequence into two sequences of length $N/2$ whose DFTs we denote by $V_0(k)$ and $V_1(k)$ respectively.

The synthesis side of the system is shown in Figure 3(b). Here, we upsample the sequences $V_0(k)$ and $V_1(k)$ as explained in Section 2.2. This gives us two sequences $U_0(k)$ and $U_1(k)$ of length N . These are then multiplied pointwise with $F_0(k)$ and $F_1(k)$ which are the DFTs of N length sequences $f_0(n)$ and $f_1(n)$ respectively. They are the synthesis filters corresponding to the analysis filters $h_0(n)$ and $h_1(n)$ respectively. The output of the synthesis filter bank is thus given by $X_R(k) = F_0(k)U_0(k) + F_1(k)U_1(k)$. The reconstructed signal is then the inverse DFT (IDFT) of $X_R(k)$, namely, $x_R(n)$.

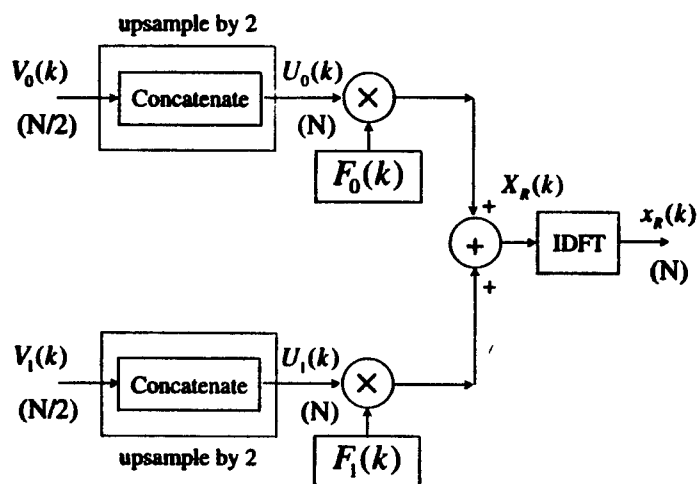
Note that, the PRCC filter bank is a framework based in the discrete frequency and is different from the work proposed by several authors before on fast implementation of FIR filter banks based on the FFT [6]. We now present the conditions for perfect reconstruction for the PRCC banks in the next subsection.

2.4 Conditions for Perfect Reconstruction

Since the input sequence is of length N it follows from the previous subsection, that the analysis and the synthesis bank filters also have a support of N samples. Furthermore, we assume that N is even and



(a)



(b)

Figure 3: PRCC filter bank system. (a) analysis filter bank and (b) synthesis filter bank.

all the sequences and filters are real valued. The conditions for perfect reconstruction in this case are obtained in a manner similar to that described in [7]. They are cyclic counterparts of the corresponding linear relationships. The transfer function of the analysis-synthesis system is given by

$$\begin{aligned}\hat{X}(e^{j2\pi k/N}) &= \frac{1}{2}[H_0(e^{j2\pi k/N})F_0(e^{j2\pi k/N}) + H_1(e^{j2\pi k/N})F_1(e^{j2\pi k/N})]X(e^{j2\pi k/N}) \\ &= \frac{1}{2}[H_0(-e^{j2\pi k/N})F_0(e^{j2\pi k/N}) + H_1(-e^{j2\pi k/N})F_1(e^{j2\pi k/N})]X(-e^{j2\pi k/N})\end{aligned}\quad (17)$$

for $k = 0, 1, \dots, N-1$. The second term in (17) is the alias term. To cancel this term we choose

$$F_0(e^{j2\pi k/N}) = H_1(-e^{j2\pi k/N}) \quad (18)$$

$$F_1(e^{j2\pi k/N}) = -H_0(-e^{j2\pi k/N}) \quad (19)$$

In addition, substituting (18) and (19) into (17) gives

$$\hat{X}(e^{j2\pi k/N}) = \frac{1}{2}[H_0(e^{j2\pi k/N})H_1(-e^{j2\pi k/N}) - H_1(e^{j2\pi k/N})H_0(-e^{j2\pi k/N})]X(e^{j2\pi k/N}) \quad (20)$$

Thus, for perfect reconstruction we require

$$H_0(e^{j2\pi k/N})H_1(-e^{j2\pi k/N}) - H_1(e^{j2\pi k/N})H_0(-e^{j2\pi k/N}) = ce^{-j2\pi L/N} \quad (21)$$

$k = 0, 1, \dots, N-1$. Here c is a real constant and L corresponds to a circular shift.

Following [7] we require that $H_0(e^{j2\pi k/N})$ satisfy the power complementarity condition given by

$$|H_0(e^{j2\pi k/N})|^2 + |H_0(-e^{j2\pi k/N})|^2 = 2 \quad (22)$$

$k = 0, 1, \dots, N-1$. In other words we require that $|H_0(e^{j2\pi k/N})|^2$ be a *half band filter* or an *equiripple filter* that is, antisymmetric about $\omega = \pi/2$ or equivalently $n = N/4$. For perfect reconstruction we choose

$$H_1(e^{j2\pi k/N}) = -e^{j2\pi k/N} H_0(-e^{-j2\pi k/N}) \quad (23)$$

Using (23) in (18) and (19) we get

$$F_0(e^{j2\pi k/N}) = e^{j2\pi k/N} H_0(e^{-j2\pi k/N}) \quad (24)$$

$$F_1(e^{j2\pi k/N}) = e^{j2\pi k/N} H_1(e^{-j2\pi k/N}) \quad (25)$$

Equations (22), (23), (24) and (25) ensure that the filters satisfy the equivalent of paraunitary conditions in this domain [1, 7] and hence satisfy the cyclic orthogonality relationships given by

$$\sum_{n=0}^{N-1} h_i(n)h_j((n+2\ell) \bmod N) = \delta(i, j)\delta(2\ell \bmod N, 0) \quad (26)$$

where $\ell \in \mathcal{Z}$, the set of integers and $i, j = 0, 1$. Note that these are the cyclic equivalents of similar relationships satisfied by orthogonal or paraunitary filter banks [7].

2.5 Procedure for Constructing PRCC Filter Banks

To obtain the filters $H_1(e^{j2\pi k/N})$, $F_0(e^{j2\pi k/N})$ and $F_1(e^{j2\pi k/N})$ we first need to design the prototype filter $H_0(e^{j2\pi k/N})$. For this we require the half band filter $H(e^{j2\pi k/N})$ defined as

$$H(e^{j2\pi k/N}) = H_0(e^{j2\pi k/N})H_0(e^{-j2\pi k/N}) \quad (27)$$

$k = 0, 1, \dots, N-1$. Note that $H(e^{j2\pi k/N})$ is a zero phase filter. Thus, the design of $H_1(e^{j2\pi k/N})$ or $H(k)$ is equivalent to assigning a value to each DFT coefficient as follows. Assuming $0 \leq H(k) \leq 1$, for some $H(k)$

1. $H(N-k) = H(k)$
2. $H(N/2-k) = 1 - H(k)$
3. $H(N/2+k) = H(N/2-k)$

$H_0(k)$ can now be designed by taking into account the fact that

$$H(e^{j2\pi k/N}) = |H_0(e^{j2\pi k/N})|^2 \quad (28)$$

Therefore,

$$|H_0(e^{j2\pi k/N})| = H(e^{j2\pi k/N})^{1/2} \quad (29)$$

Given that in general, $H_0(e^{j2\pi k/N})$ has the form

$$H_0(e^{j2\pi k/N}) = |H_0(e^{j2\pi k/N})| e^{j\phi(k)} \quad (30)$$

we can now add the phase term $\phi(k)$. Since we require that $h_0(n)$ be real, $\phi(k)$ is antisymmetric about $N/2$. The filters $H_1(k)$, $F_0(k)$ and $F_1(k)$ can now be derived using the relations (23), (24) and (25).

Example. $N = 8$.

Let $H(0) = 0.75$. Then $H(4) = 0.25$

Let $H(1) = 0.37$. Then $H(7) = 0.37$

and $H(3) = H(5) = 0.63$

Finally, $H(2) = 1 - H(2) = H(6) = 0.5$.

Thus,

$H(k) = \{0.75, 0.37, 0.5, 0.63, 0.25, 0.63, 0.5, 0.37\}$, and hence

$h(n) = \{0.5, 0.0166, 0.0, 0.1084, 0.0, 0.1084, 0.0, 0.0166\}$

Note that the non-zero even indexed points of $h(n)$ have value 0.

From (29), this gives us $|H_0(k)| = \{0.866, 0.6082, 0.7071, 0.7937, 0.5, 0.7937, 0.7071, 0.6082\}$

Let us choose

$\phi(k) = \{0, 1.1168, 0.2302, -2.6746, 0.0, 2.6746, -0.2302, -1.1168\}$

This gives $h_0(n) = \{0.2355, 0.2746, 0.2215, -0.1162, 0.4565, -0.13721, -0.2305, 0.1618\}$

This completes the design of $H_0(k)$. The filters $H_1(k)$, $F_0(k)$ and $F_1(k)$ can now be determined. Note the flexibility and ease of design that this method offers for designing filters to specification.

2.6 Frequency Sampled Implementation of Bandlimited DWT

The power spectrum of the Meyer scaling function $|\Phi(\omega)|^2$, satisfies the conditions of an interpolating filter in that it is antisymmetric about $\omega = \pi$. It follows that if it is properly sampled in the frequency domain then these properties can be retained and a half band filter can be obtained. To determine the rate at which it needs to be sampled, it is important to note that the function needs to be sampled symmetrically about the angular frequency of π units. This means, if the required filter size is N samples, where N is assumed even, then the samples should be $\Delta\omega = 4\pi/N$ units apart in frequency so that the samples are antisymmetric about $N/4$. The square root of these samples gives the samples of the low pass filter $H_0(k)$, which can now be used in the PRCC framework. The shape of the low pass and high pass filter thus obtained are shown in Figure 4.

3 Continuous-dilation discrete-time self-similar signals and linear scale-invariant systems

This research addresses the problem of defining and representing discrete-time self-similar signals and systems. The study of the discrete-time self-similar processes was motivated in part by the previous work of Wornell and colleagues [10, 11, 9] in continuous time. They provide formulations involving continuous-time, scale-invariant signals and systems. They also provide a detailed study of such systems for dyadic scale factors. Our research provides answers to questions such as: Is it possible to define purely discrete-time, self-similar signals? Are there formulations of discrete-time, scale-invariant systems? How do we provide a definition of dilation or scaling of discrete-time signal that is general enough to provide non-trivial self-similar signals and scale-invariant systems? The answer to the third question holds the key to answering the first two questions. A key result of the research is the demonstration of the fact that it is possible to define scaling or dilation in such a way that is continuous even though the signal itself is discrete-time. Hereafter, we will use the term scaling exclusively to mean dilation. Using this definition of scaling, we develop definitions and constructions of deterministic and stochastic, discrete-time, self-similar signals and discrete-time scale-invariant systems.

3.1 Scaling in Discrete-Time

3.1.1 Discrete-Time Scaling Operation

Generally the scaling or dilation operation of a discrete-time signal $x(n)$ by an arbitrary factor is not well defined. It is difficult to obtain an interpretation of scaling in the discrete-time domain that is as unambiguous as that in the continuous-time domain. Operations such as upsampling, interpolation, downsampling and fractional sampling rate alteration [7] can have a scaling interpretation. However, such operations cannot handle scaling factors over a continuum. We present here a different approach to discrete-time scaling that can handle continuous scaling factors. We define the discrete-time scaling

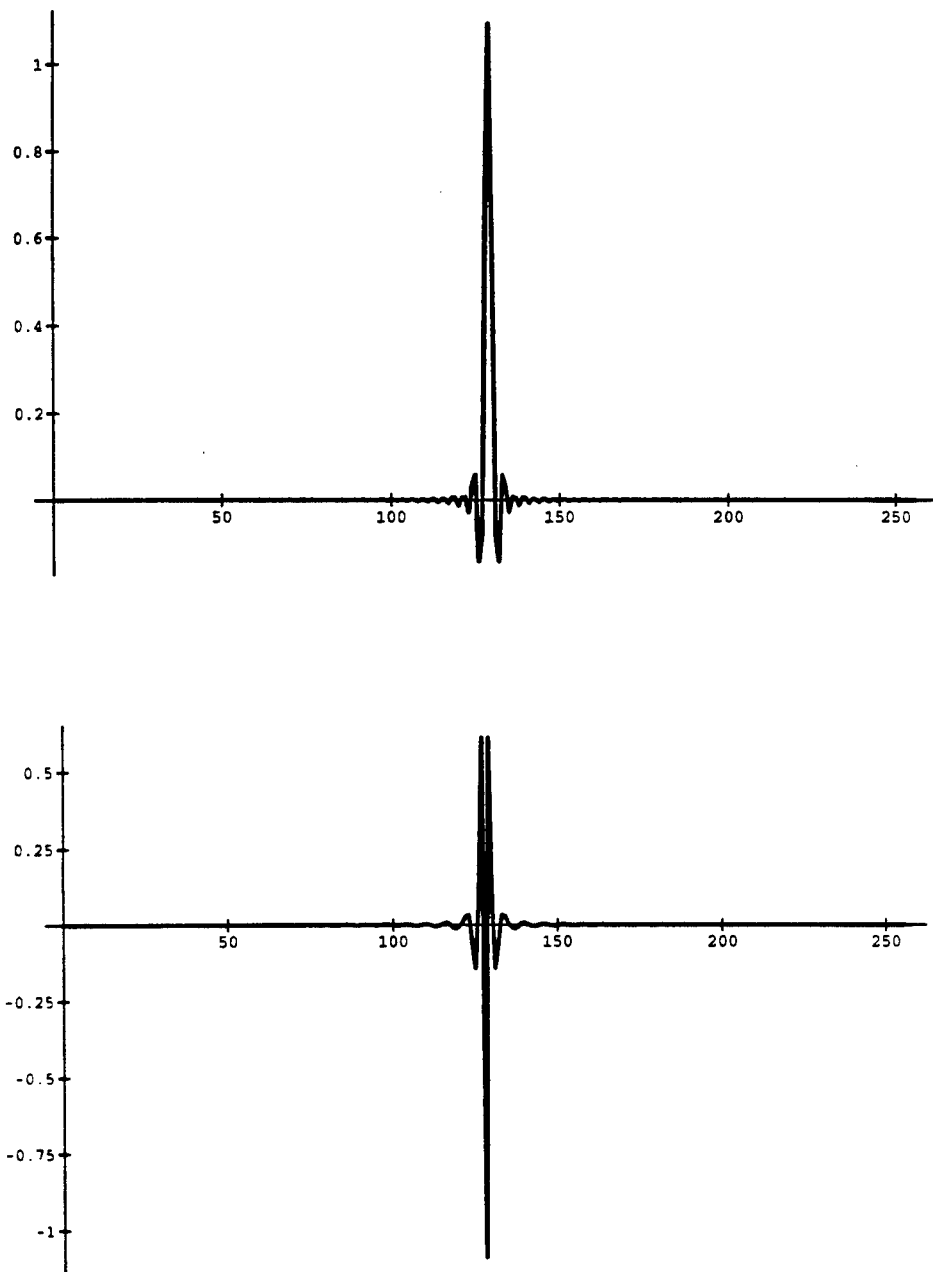


Figure 4: Low pass and high pass filter response obtained by frequency sampling the power spectrum of the Meyer scaling function

operation in a way that effectively amounts to converting $x(n)$ into a continuous-time signal through an invertible mapping, applying the scaling operation to the continuous-time signal and finally inverse mapping the signal back to the discrete-time domain. The actual definition is based on operations in the frequency domain.

Let $f(t)$ be a continuous-time signal and $F(\Omega)$ its Fourier transform:

$$F(\Omega) = \mathcal{F}\{f(t)\} = \int_{-\infty}^{+\infty} f(t)e^{-j\Omega t} dt, \quad (31)$$

where $-\infty < \Omega < +\infty$. If $f(t)$ is scaled by a ($a > 0$), its Fourier transform becomes

$$\mathcal{F}\{f(t/a)\} = aF(a\Omega), \quad -\infty < \Omega < +\infty. \quad (32)$$

Thus, for a continuous-time signal, a scaling in time can be accomplished in principle by a frequency-scaling of its Fourier transform in the opposite direction along with an amplitude scaling. Now, consider a discrete-time sequence $x(n)$ whose Fourier transform is

$$X(\omega) \equiv \mathcal{G}\{x(n)\} = \sum_n x(n)e^{-j\omega n}. \quad (33)$$

The function $X(\omega)$ is 2π -periodic. If we try to define a discrete-time scaling operation by adapting (32) to (33), it will only work for integer values of a because of the 2π -periodicity requirement on the Fourier transform of a discrete-time signal. This corresponds to upsampling the discrete-time signal by an integer factor of a . The implementation of our discrete-time continuous scaling operation is as follows (see Figure 5).

1. Given is a discrete-time signal $x(n)$ with (the 2π -periodic) Fourier transform $X(\omega)$.
2. Map the principal interval $\omega \in [-\pi, \pi]$ to continuous frequency Ω (the real line) through an invertible transformation $\Omega = f(\omega)$.
3. Dilate $Y(\Omega) \equiv X(f^{-1}(\Omega))$ by the required dilation factor a to form $Y_a(\Omega) \equiv aY(a\Omega)$.
4. Form $X_a(\omega) = Y_a(f[\omega])$
5. The sequence $x_a(n)$ resulting from the inverse Fourier transformation of $X_a(\omega)$ is the continuous dilation of $x(n)$ by a

$$x_a(n) = a\mathcal{G}^{-1}\{X[f^{-1}(af(\omega))]\}. \quad (34)$$

where \mathcal{G}^{-1} denotes inverse discrete-time Fourier transform.

Some examples of $f(\omega)$ ($\omega \in [-\pi, \pi]$) are:

- *Bilinear transform.* $\Omega = f(\omega) = 2 \tan(\omega/2)$.
- *1/ ω -based transform.* $\Omega = f(\omega) = \frac{\omega}{\pi - |\omega|}$.
- *log-based transform.* $\Omega = f(\omega) = \text{sgn}(\omega) \ln \left(\frac{\pi}{\pi - |\omega|} \right)$.

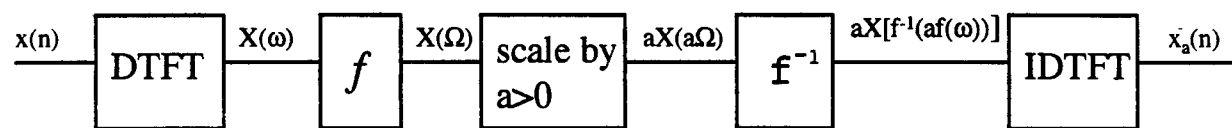


Figure 5: Block diagram of the discrete-time, continuous scaling operator.

3.1.2 System Properties

Let \mathcal{S}_a denote the discrete-time scaling operator defined above. It is straightforward to verify that \mathcal{S}_a has the following properties:

1. \mathcal{S}_a is a linear operator.
2. \mathcal{S}_a ($a \neq 1$) is a time-varying operator.
3. $\mathcal{S}_1\{x(n)\} = x(n)$ as expected. This corresponds to the non-scaling case.
4. The inverse operator $\mathcal{S}_a^{-1}\{x(n)\} = \mathcal{S}_{1/a}\{x(n)\}$ is discrete-time scaling operation with parameter $1/a$.
5. Commutativity

$$\mathcal{S}_a\{\mathcal{S}_b\{x(n)\}\} = \mathcal{S}_b\{\mathcal{S}_a\{x(n)\}\} = \mathcal{S}_{ab}\{x(n)\} \quad (35)$$

6. If the discrete-time Fourier spectrum in the principal interval $[-\pi, \pi]$ of an input discrete-time signal is a function of $f(\omega)$, i.e.,

$$X(\omega) = T[f(\omega)], \quad (36)$$

and the function $T(\omega')$ satisfies

$$T(a\omega') = C(a)T(\omega'), \quad (37)$$

where $C(a)$ is a function of a , then the output of the discrete-time scaling operator is

$$\mathcal{S}_a\{x(n)\} = aC(a)x(n). \quad (38)$$

Property 6 provides some interesting insights into the discrete-time scaling operation. It implies that if the inverse Fourier transform of the function $T[f(\omega)]$ exists, the corresponding time sequence represents an eigen-function of the system. Also, when the input spectrum satisfies (36) and (37), for example,

$$T(\omega') = \omega'^r \text{ and hence } X(\omega) = T[f(\omega)] = [f(\omega)]^r, \quad (39)$$

the output spectrum is identical to the input within an amplitude factor $aC(a)$ (a^{r+1} in the example). In other words, the signal is identical to a scaled version of itself within an amplitude factor.

3.2 Discrete-time Self-Similar Signals

3.2.1 Self-Similarity

Two types of self-similar signals will be discussed in here: deterministic and stochastic.

Definition: A discrete-time sequence $x(n)$ is deterministically self-similar or homogeneous with degree H if it satisfies the following relation.

$$\mathcal{S}_a\{x(n)\} = a^{-H}x(n) \quad (40)$$

for any $a > 0$. A random process $X(n)$ is said to be statistically self-similar with degree H if it satisfies the following equation

$$\mathcal{S}_{a,a}\{R_X(n, n')\} = a^{-2H} R_X(n, n') \quad (41)$$

for any $a > 0$, where $R_X(n, n')$ denotes the auto-correlation function of sequence $X(n)$, and $\mathcal{S}_{a,b}\{x(m, n)\}$ for a 2-D function $x(m, n)$ is defined in lines similar to that of \mathcal{S}_a . However, the scaling operation is applied on both m and n dimensions.

3.2.2 Discrete-Time Homogeneous Signal

As mentioned in section 3.1.2, the time sequence corresponding to inverse Fourier transform of function $[f(\omega)]^r$, if exists, satisfies (40) with $H = -(r + 1)$. Thus, by choosing a function $[f(\omega)]^r$ which is absolutely integrable in $-\pi$ to π , we can derive a class of discrete-time homogeneous functions. This class of homogeneous functions could provide a model for discrete-time self-similar process in practice. They also serve as eigen-functions of the discrete-time scaling operator previously defined.

As we know, the class of continuous-time, regular, homogeneous functions such as $f(t) = 1$ is limited. Truly continuous homogeneous signals corresponding to the spectrum Ω^r do not exist because it is not a valid Fourier spectrum. In our formulation of discrete-time self-similar functions, we are able to derive purely discrete time sequences as long as $[f(\omega)]^r$ defines a valid discrete-time Fourier spectrum. Non-trivial discrete-time homogeneous functions actually exist and can be derived in the following ranges of r parameter with respect to different mappings.

- Bilinear Transform. $-1 < r < 1$.
- $1/\omega$ Based Transform. $-1 < r < 1$.
- \log -Based Transform. $r \neq -1, -2, -3, \dots$

Figure 6 shows some examples of discrete-time deterministic self-similar functions which are derived from discrete-time Fourier spectrum $[f(\omega)]^{1/2}$. $f(\omega)$ is chosen as bilinear transform, $1/\omega$ -based and \log -based transform respectively.

3.3 Discrete-Time Linear Scale-Invariant Systems And Self-Similar Functions

3.3.1 Discrete-Time Linear Scale-Invariant System

A linear scale-invariant (LSI) system is a linear system whose output is invariant to the scale changes of the input. Let $x(n)$, $y(n)$, ($n \in (-\infty, \infty)$) be the input and output sequence. Take an arbitrary discrete-time sequence $h(k)$ ($k = 1, 2, \dots, K$) and let

$$y(n) = \sum_{k=1}^K h(k) \mathcal{S}_k\{x(n)\}/k. \quad (42)$$

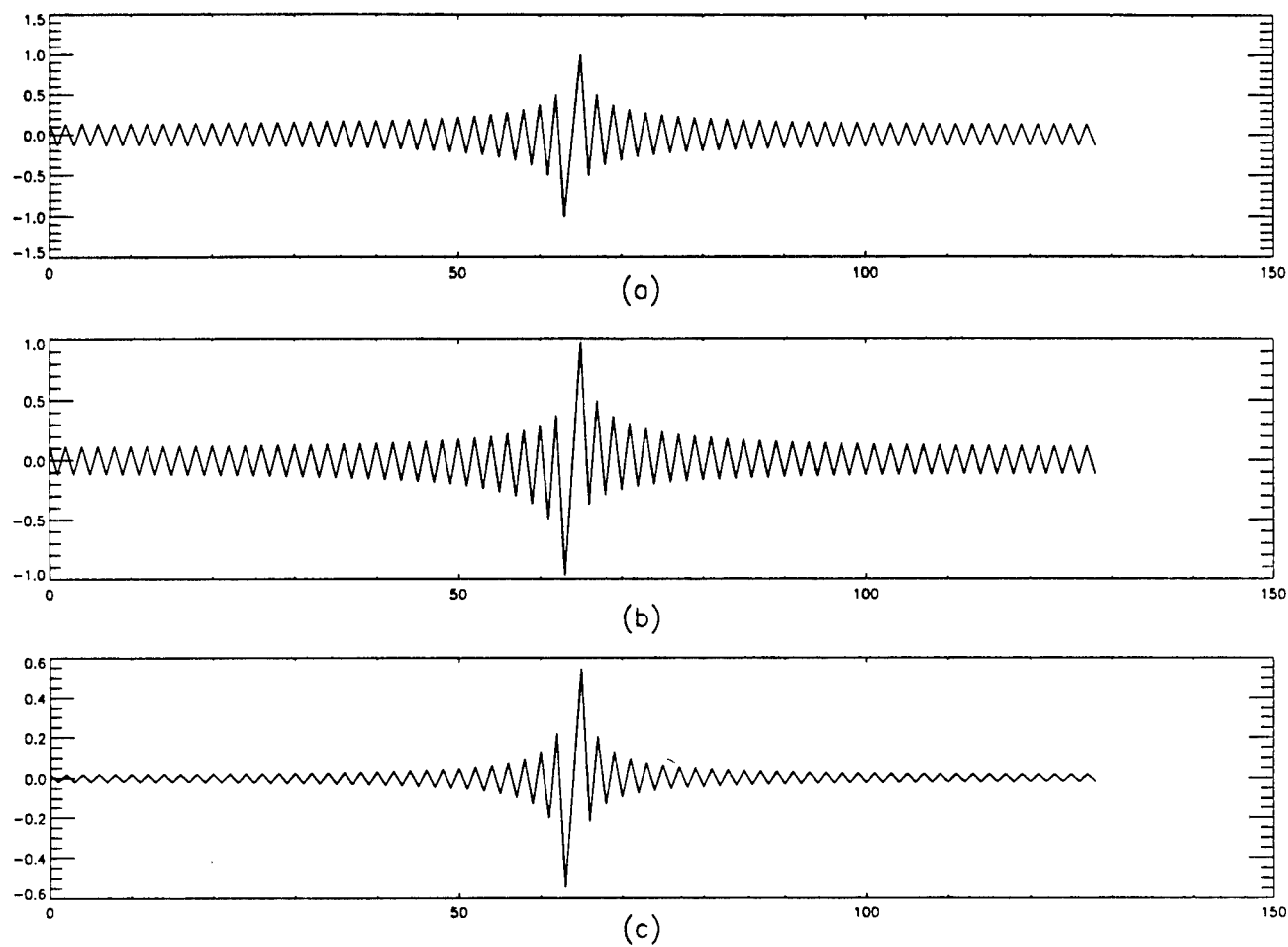


Figure 6: Eigen-functions of the discrete-time scaling operation system with respect to (a) bilinear transform (b) $1/\omega$ -based transform (c) \log -based transform when $r = 0.5$. The imaginary parts of the functions are shown.

This defines a discrete-time LSI system. The output of the system is the summation of a series of dilated (by k) input sequence, weighted by the kernel $h(k)$. Figure 7 shows the procedure for the implementation of the system. Note that the choice of the one dimensional kernel $h(k)$ is arbitrary. Also, as mentioned in section 3.2, the discrete-time sequences corresponding to spectrum $[f(\omega)]^r$ are eigen-functions of the discrete-time scaling operator. Inherently, they also serve as the eigen-functions of the discrete-time LSI system.

3.3.2 Discrete-time Statistically Self-Similar Signal

As mentioned in [5, 10], most physical processes that exhibit statistical self-similarity are fundamentally non-stationary. The statistical properties of the signal change with time, but remain invariant with time scale. In this section we provide a model for such non-stationary self-similar random processes using the discrete-time LSI system. Our implementation in discrete-time domain is based on the following property of the discrete-time LSI system.

Theorem: If the input sequence of a discrete-time LSI system is discrete-time zero-mean white noise, the output sequence of the system is non-stationary and statistically self-similar which satisfies condition (41) with $H = -1$.

proof: See [12].

Hence we can construct a non-stationary self-similar random signal with parameter $H = -1$ by passing a discrete-time zero-mean white noise through the discrete-time LSI system. By passing the signal thus obtained through the system again, a non-stationary self-similar random signal with parameter $H = -2$ is then acquired. Following this scenario, we are able to formulate a non-stationary self-similar random process with parameter H being an arbitrary negative integer. Note that the choice of the one dimensional kernel $h(k)$ in our discrete-time LSI system is essentially arbitrary. We can choose a specific kernel $h(k)$ so that the output of the system exhibits the properties of the studied physical self-similar processes. As there is no restriction on the length of the kernel, a rich class of existing FIR or IIR filters can be applied to model the behavior of a large variety of self-similar random processes in practice.

Figure 8 demonstrates the effect of passing a discrete-time zero-mean white noise through a discrete-time LSI system. The output is a discrete-time stochastic self-similar signal. As is known, if the system output is wide-sense stationary, the 2-D plot of auto-correlation function of the output signal will consist of a series of diagonal straight contour. The auto-correlation plot in Figure 8 clearly demonstrate the non-stationary property of the output signal.

4 Scaling Mixtures in Multichannel Data

The detection and classification of objects in images is obviously not restricted to visual imagery. It can arise in other contexts as well. In particular, our investigation examined the role of scaling in multichannel recordings for bearing and velocity estimation of wideband plane waves which would have

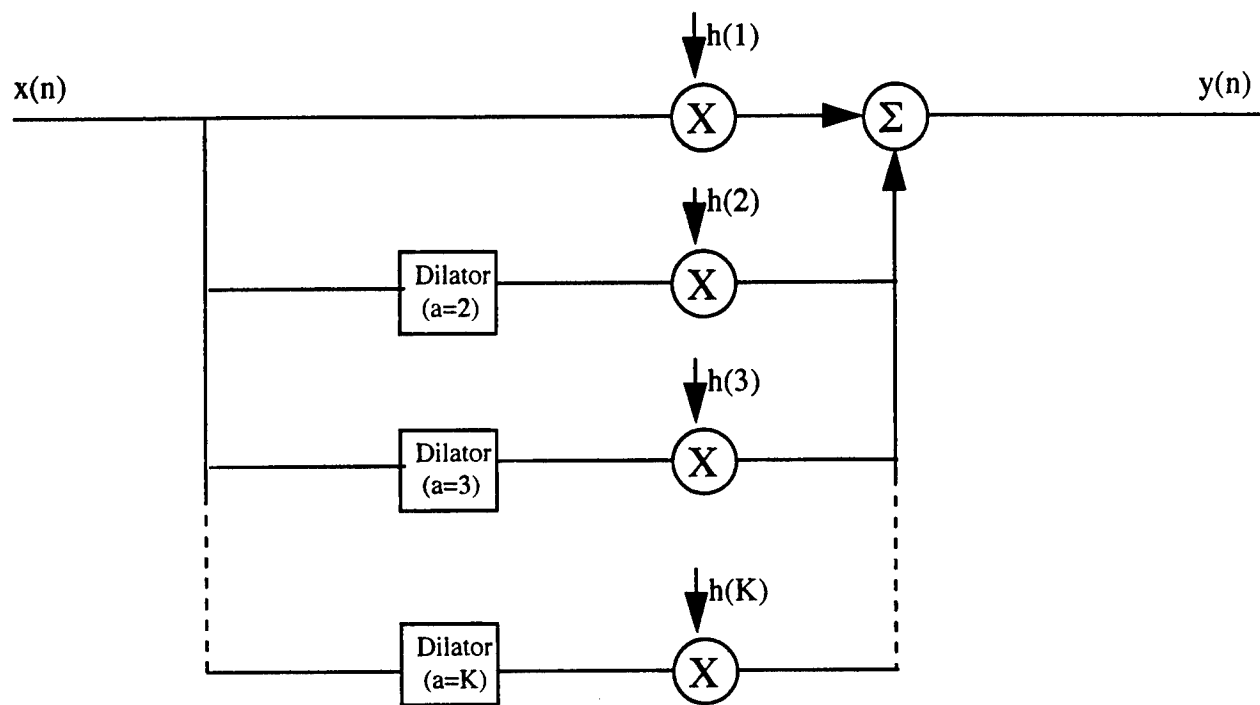


Figure 7: System diagram of the discrete-time LSI system

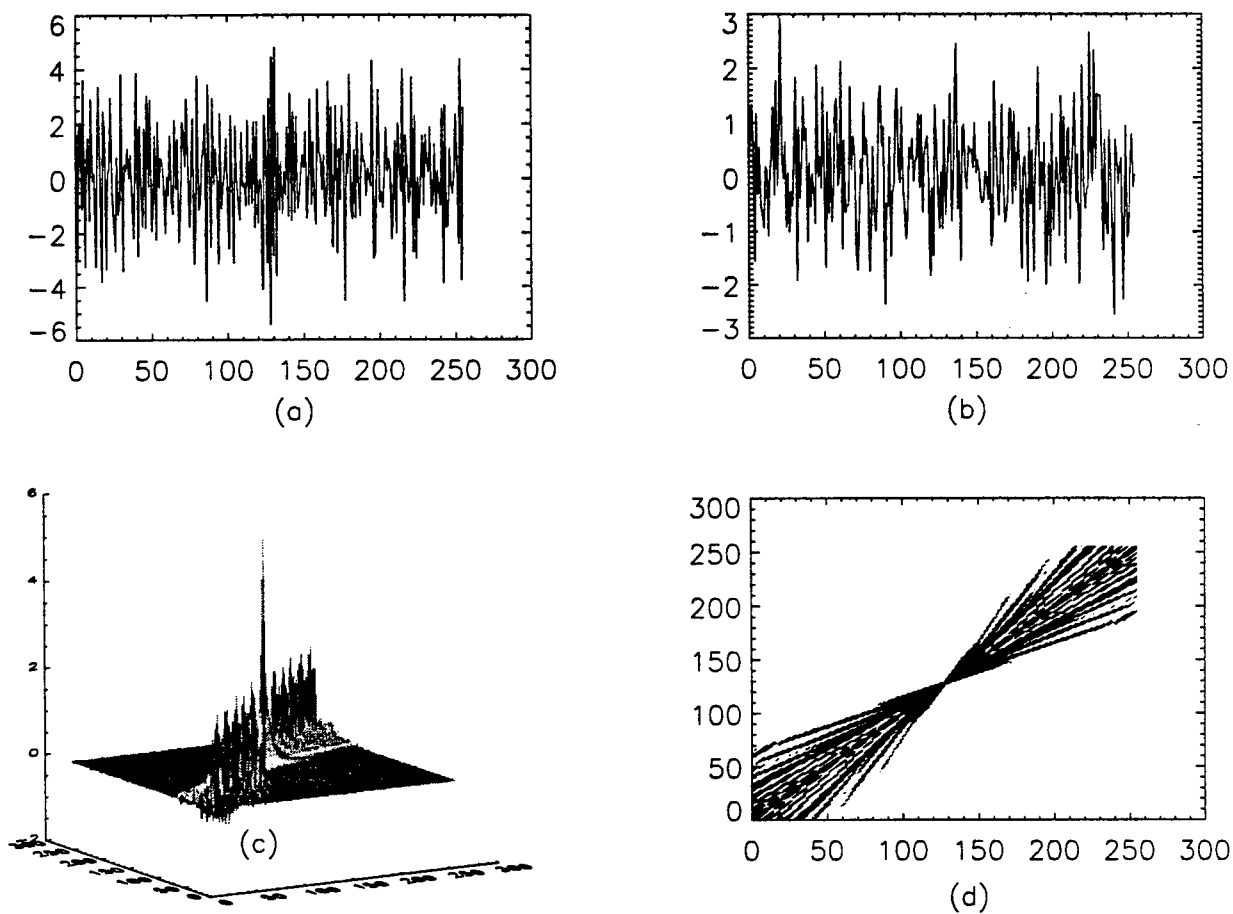


Figure 8: Simulation of passing a zero-mean white noise through discrete-time LSI system. The 6-tap Hanning window is used as 1-d kernel for the discrete-time LSI system. (a) system input (b) system output (c) auto-correlation function of the output (d) contour plot of the auto-correlation function of the output

potential application to radar and sonar.

4.1 Scaling Mixtures in Bearing Estimation

Suppose we have the two-dimensional plane wave

$$s(\mathbf{u}, t) = Af(t - \mathbf{k} \cdot \mathbf{u}/c) \quad (43)$$

impinging on a uniform linear array of N receiving elements located at $(x_1, 0), (x_2, 0), \dots, (x_N, 0)$ respectively, where $f(x)$ is a square-integrable function, $\mathbf{u} = [u_x, u_y]$ is the spatial position vector, $\mathbf{k} = [k_x, k_y]$ is the direction vector, c is the speed of propagation of the plane wave in the medium and $x_n = (n-1)d$ with d being the sensor spacing. In the narrowband formulation, $f(x)$ is either a complex exponential or something close to it. However, the assumption of $f(x)$ possessing finite energy rules out such a model here. The recording in the sensors is denoted $x_n(t)$ for $n = 1, \dots, N$. We have

$$x_n(t) = Af\left(t - \frac{k_x d}{c}(n-1)\right) + n(t) \quad (44)$$

where $n(t)$ is the additive noise which could be due to any or all of clutter, sensor noise and jamming. Examining (44) we find that a snapshot in time of the sensor recordings consists of samples of the function $f(x)$ taken at $x = t - \frac{k_x d}{c}(n-1)$ for $n = 1, \dots, N$. An integration over time of the correlation matrix of snapshots yields (ignoring the noise for the time being) the $N \times N$ Toeplitz matrix

$$\mathbf{R} = A^2 \begin{bmatrix} R_{k_x}(0) & R_{k_x}(1) & \dots & R_{k_x}(N-1) \\ R_{k_x}(1) & R_{k_x}(0) & \dots & R_{k_x}(N-2) \\ \vdots & \vdots & \ddots & \vdots \\ R_{k_x}(N-1) & R_{k_x}(N-2) & \dots & R_{k_x}(0) \end{bmatrix} \quad (45)$$

where

$$R_{k_x}(\tau) = R_f\left(\frac{k_x d}{c}(\tau)\right) \quad (46)$$

with

$$R_f(\tau) \equiv \int f(t)f(t+\tau) dt \quad (47)$$

being the temporal autocorrelation of the function $f(t)$. Thus, we find that the angle of arrival as determined by k_x , reveals itself as a temporal scaling $k_x d/c$ in the autocorrelation matrix. The angle of arrival determination problem is thus transformed into one of determining the dilation or scaling from the spatial autocorrelation matrix in the presence of noise. With multiple plane waves, the spatial autocorrelation matrix consists of the sum of autocorrelations at different scalings and hence constitutes a scaling mixture.

4.2 Scaling Mixtures in Velocity Estimation

Suppose the pulse $f(t)$ above is reflected off a moving object whose velocity component along the direction of travel of the pulse is v . Then, upon reflection, the pulse $f(t)$ suffers a time scaling of

$(c + v)/c$, that is, if $f(t)$ is symmetric, the received pulse is a time-delay of $f(\frac{c \pm v}{c}t)$. A temporal autocorrelation of the received pulse would result in a time-scaled autocorrelation of the transmit pulse and would be independent of the time-delay. Clearly, the time-delay is an estimate of the range while the scaling provides an estimate of the velocity v . The temporal autocorrelation when there are multiple objects at different velocities again results in a scaling mixture of the autocorrelation of the pulse.

4.3 Scale Estimation

In the particular case that the function $f(t)$ is a wavelet then so is its autocorrelation. A wavelet transform of the autocorrelation scaling mixture using this wavelet would lead to scale determination since the scale coordinates of peak positions in the wavelet transform correspond to the scale parameters of the individual components in the input [4]. However, it is likely that high resolution estimation of the scaling parameters would require procedures analogous to the subspace methods developed for the narrowband case. We have begun investigating the development of such methods. This is a promising area for further research.

Acknowledgement

The Principal Investigator wishes to thank graduate students Ajit Bopardikar, Greg Pettis and Wei Zhao for collaborating on the research results documented here, Dr. James Michels, Rome Laboratory and Dr. Raman Unnikrishnan, Head of the Department of Electrical Engineering, RIT, for their support.

References

- [1] A. S. Bopardikar. Speech encryption using wavelet packets. Master's thesis, Indian Institute of Science, Bangalore, India, 1995.
- [2] A. S. Bopardikar, M. R. Raghuveer, and B. S. Adiga. Perfect reconstruction circular convolution filter banks and their application to the implementation of bandlimited discrete wavelet transform. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3665-3668, Munich, Germany, April 1997.
- [3] A. S. Bopardikar, M. R. Raghuveer, and B. S. Adiga. PRCC filter banks: Theory, implementation and application. In *Proceedings of SPIE, Wavelet Applications in Signal and Image Processing V*, San Diego, CA, USA, August 1997.
- [4] J. O. Chapa. *Matched wavelet construction and its application to target detection*. PhD thesis, Rochester Institute of Technology, 1995.
- [5] J. Feder. *Fractals*. Plenum Press, New York, 1988.

- [6] O. Rioul and P. Duhamel. Fast algorithms for discrete and continuous wavelet transforms. *IEEE Trans. On Information Theory*, 38(2), March. 1992.
- [7] P. P. Vaidyanathan. *Multirate systems and filter banks*. Prentice-Hall, Englewood Cliffs, New Jersey, 1992.
- [8] P. P. Vaidyanathan and A. Kirac. Theory of cyclic filter banks. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 2449–2452, Munich, Germany, April 1997.
- [9] G. W. Wornell. Emerging applications of multirate signal processing in digital communications. *Proceedings of IEEE*, 84(4), April 1996.
- [10] G. W. Wornell. *Signal Processing with Fractals*. Prentice-Hall, Upper Saddle River, New Jersey, 1996.
- [11] G. W. Wornell and A. V. Oppenheim. Wavelet-based representations for a class of self-similar signals with application to fractal modulation. *IEEE Trans. on Information Theory*, 38(2):785–800, Mar. 1992.
- [12] W. Zhao and R.M. Rao. Discrete-time self-similar signals and scale-invariant systems. Technical Report TR-Zhrao-97-01, Center for Imaging Science, Rochester Institute of Technology, Rochester, NY, 14623 USA. October 1997.

IPL HTML Interface Performance Evaluation

Scott Spetka
Associate Professor
Department of Computer Science

State University of New York
Institute of Technology at Utica/Rome
Route 12 North
Utica, New York 13504

Final Report for:
Summer Research Extension Program
Rome Laboratory

Sponsored by:
Air Force Office of Scientific Research
Bolling Air Force Base, Washington, D.C.

December 1997

IPL HTML Interface Performance Evaluation

Scott Spetka
Assistant Professor
Department of Computer Science
State University of New York
Institute of Technology at Utica/Rome

Abstract

This grant developed tools for use in performance evaluation and benchmarking of the IPL HTML interface to multimedia databases. The major contributions of the grant, described in this report, were in the areas of developing benchmarking support for IPL HTML, developing benchmarking tools for use in ongoing performance evaluations of the evolving IPL HTML software, and experimentation with software that could be used in the IPL HTML environment to enhance performance. Each of these contributions is described further in the report. The report concludes with a summary of the contributions to the IPL HTML system that were developed under this grant.

IPL HTML Interface Performance Evaluation

Scott Spetka

Introduction

This grant developed tools for use in performance evaluation and benchmarking of the IPL HTML interface to multimedia databases. The major contributions of the grant, described in this report, were in the areas of developing benchmarking support for IPL HTML, developing benchmarking tools for use in ongoing performance evaluations of the evolving IPL HTML software, and experimentation with software that could be used in the IPL HTML environment to enhance performance. Each of these contributions is described further in the report. The report concludes with a summary of the contributions to the IPL HTML system that were developed under this grant.

Benchmarking Support

Benchmarking tools were integrated into the implementation of IPL HTML to enhance the SQL query interface and the Harvest query interface with real-time feedback. This code allowed experimentation with queries of varying complexity. The output displays detailed system performance information. It includes numbers of page faults, cpu usage, numbers of I/O operations and elapsed time. The most useful metric proved to be the elapsed time since internal performance metrics were not available for Harvest and Sybase. Experiments using these benchmarking tools proved that the systems were comparable in speed with Harvest outperforming Sybase for some types of queries.

The experiment proved that Harvest could be an important tool for an information analyst. It is particularly effective in cases where queries are very selective (produce small results compared to the data space searched). Harvest does not allow "Less than" or "greater than" operators in queries and does not support complex aggregate operations, like averages, supported by general purpose database systems like Sybase. The study confirmed the relevance of Harvest to the continuing development of the IPL HTML architecture. The benchmarking tools were implemented in the actual IPL HTML code where they can be used for continual evaluation of system performance.

Benchmarking Tools

IPL HTML is implemented as an extension to a WWW server. Measuring the performance of CGI-BIN extensions to the HTTPD requires tools to generate a load on the server. The study considered the use of WebStone, a tool that automatically generates a load on a server by submitting a mix of queries. Experiments with this tool demonstrated its usefulness for IPL HTML. Experiments were conducted at SUNY Institute of Technology using university equipment made available for this research.

The WebStone software was installed on the SUNY Institute of Technology DogNET Unix network. Supporting software for WebStone was made and installed. This included Perl 5.0 and Gnu Make. This included Gcc V 2.7.2. WebStone was then installed. Make was used to develop other software as well.

System were reconfigured and WWW pages were added to the SUNY Tech WWW server to keep track of webserver traffic. An account was added for WebStone on Growl, a Sun Solaris system. The following statistical software was installed to monitor server load: Webstat 2.0, GWstat 2.3, 3Dstat 2.1, and Analog 2.0. Determining when a system must be enhanced to support increased loads is an important part of performance evaluation.

Setting up and testing WebStone turned out to be difficult. WebStone did not run on FreeBSD, the operating system in use on most of the DogNET machines. It compiled but did not work right. The documentation is very poor and there is not much of it. Reddog is running Solaris 2.5.1 but it is the only system running Solaris. WebStone needs at last two computers to run, a master and a client. Spot and Growl both run Sun OS 4.1.3. Gcc needed to be installed on the two SunOS machines and some other system problems needed to be fixed before some test results should be obtained from WebStone.

Performance Enhancing Tools

Performance Enhancing Tools can be used to improve performance of the IPL HTML server. The system is implemented as an extension of a standard WWW server so that any tools that enhance WWW server performance would be of interest for some functions of the IPL HTML server. Experimentation with software that could be used to enhance WWW server performance provides direction as to whether use of this type of tool should be considered for the IPL HTML environment. *Squid* is a system of caching software for web clients and front end software for web servers. Experiments were performed at SUNY Institute of Technology on the DogNET to evaluate whether this software should be considered for IPL HTML.

Squid was installed and configured on Pitbull. Pitbull was chosen because it had sufficient disk space and was not running any other services on DogNET. Squid Squid can be configured in two ways. In a web cache proxy configuration, squid is used as a cache for incoming web traffic from the Internet. It is best suited for a situation where a number of analysts are accessing the same remote server. The first time a page is requested from an outside server it is stored on the cache proxy. Any other requests to these pages will be obtained from the cache proxy and not from the remote server. In this configuration outside network traffic is reduced and the pages that are requested by multiple users can be delivered more quickly.

The second way that Squid can be configured is as an accelerator. In this configuration squid acts as a mirror to an existing server. One way to use this configuration is to put the webserver on the inside of a firewall and to have the accelerator on the outside of the firewall. In this configuration, Squid maintains a cache of pages from the associated WWW server and delivers them without placing a load on the production WWW server. When a page is required that is not in the cache, of course Squid must retrieve it from the WWW server. As with any cacheing scheme, locality of reference in access patterns is required for successful operation. This may well be the case for IPL HTML where recently archived multimedia objects are expected to be of greater interest than older objects.

Another reason to use Squid as an accelerator is to create a mirror. For example, the primary server could be in Spain but serving data to a significant number of locations in the United States. Squid could run on a computer in the US and request people from the US to use it. This would cut down on the bandwidth requirements for higher cost international network links. The software and documentation are both good for this shareware software.

Results

An experiment was conducted using WebStone as a tool to measure the performance of the SUNY Tech WWW server with and without Squid acceleration. The study compared the speed of our webserver and the speed of Squid using it as an accelerator to our webserver. Results presented here show that going directly to the webserver is faster than going through the Squid accelerator.

Below is a summary of tests from WebStone. The experiment was performed twice on each configuration to measure server performance with and without Squid acceleration. The Squid server was running on Pitbull and the WWW server was running on Arf as indicated in the results.

First run testing the squid server running on Pitbull:

| | |
|-------------------------------|----------------------|
| WEBSTONE 2.0 results: | |
| Total number of clients: | 2 |
| Test time: | 5 minutes |
| Server connection rate: | 8.77 connections/sec |
| Server error rate: | 0.0000 err/sec |
| Server throughput: | 1.09 Mbit/sec |
| Little's Load Factor: | 1.99 |
| Average response time: | 0.2276 sec |
| Error Level: | 0.0000 % |
| Average client throughput: | 562 Kbit/sec |
| Sum of client response times: | 598.794322 sec |
| Total number of pages read: | 2631 |

Second run testing the squid server on Pitbull

| | |
|-------------------------------|----------------------|
| WEBSTONE 2.0 results: | |
| Total number of clients: | 2 |
| Test time: | 5 minutes |
| Server connection rate: | 8.87 connections/sec |
| Server error rate: | 0.0000 err/sec |
| Server throughput: | 1.03 Mbit/sec |
| Little's Load Factor: | 1.99 |
| Average response time: | 0.2248 sec |
| Error Level: | 0.0000 % |
| Average client throughput: | 531 Kbit/sec |
| Sum of client response times: | 598.99352 sec |
| Total number of pages read: | 2660 |

First run testing the www server on Arf

| | |
|-------------------------------|----------------------|
| WEBSTONE 2.0 results: | |
| Total number of clients: | 2 |
| Test time: | 5 minutes |
| Server connection rate: | 9.59 connections/sec |
| Server error rate: | 0.0000 err/sec |
| Server throughput: | 1.27 Mbit/sec |
| Little's Load Factor: | 2.02 |
| Average response time: | 0.2110 sec |
| Error Level: | 0.0000 % |
| Average client throughput: | 645 Kbit/sec |
| Sum of client response times: | 607.94653 sec |
| Total number of pages read: | 2877 |

Second run testing the www server on Arf

| | |
|-------------------------------|----------------------|
| WEBSTONE 2.0 results: | |
| Total number of clients: | 2 |
| Test time: | 5 minutes |
| Server connection rate: | 8.69 connections/sec |
| Server error rate: | 0.0000 err/sec |
| Server throughput: | 1.27 Mbit/sec |
| Little's Load Factor: | 1.99 |
| Average response time: | 0.2296 sec |
| Error Level: | 0.0000 % |
| Average client throughput: | 654 Kbit/sec |
| Sum of client response times: | 598.378061 sec |
| Total number of pages read: | 2606 |

Conclusion

The benchmarking tools developed under this grant have already proven useful in performance evaluation of the IPL HTML interface. They are part of the Evolving IPL HTML Interface. Experiments with performance enhancing tools were less conclusive. For the DogNET network, using webserver directly was faster than going to the Squid server. This could have happened for a number of reasons. It could be our network topology or differences in the configurations of systems that Squid and our WWW server are running on. Experiments showed that Squid is a stable piece of software and that further development and experimentation with Squid may be useful. In addition, experience with WebStone will prove useful for further work in IPL HTML performance evaluation. In any case, Squid would be of value if it can be run effectively as a cache proxy.

Providing an effective cacheing mechanism for IPL HTML goes beyond the basic functionality of Squid. Most of the IPL HTML functionality is implemented in CGI-BIN functions which generate pages dynamically. Data that is presented to users is derived from multimedia databases rather than stored HTML documents. The advantages of cacheing could easily be worth an effort to extend Squid to provide adequate HTML and multimedia object cacheing to allow dynamic pages to be created. It is imperative that such functionality be developed as increases in demand for Internet based services continue to outpace the increases in network bandwidth available. A cost-effective solution in the short term will be to depend on software support to improve performance and avoid expensive hardware upgrades until prices come down.

References

Squid Users Guide: <http://rk.pvt.net/doc/Squid/Welcome.html>

WebStone: <http://www.sgi.com/TEXT/Products/WebFORCE/WebStone/index.html>

Salerno, J., Spetka, et.al., "Using an HTML Interface to Integrate Heterogeneous Distributed Image Databases", IEEE Dual-Use Technologies and Applications Conference, May, 1997

Spetka, S.E., "Integrating a Multimedia Database and WWW Indexing Tools", Air Force Rome Laboratory, 1996.

Salerno, J., Spetka, S.E., Mozloom, P., Miller, R., Peck, D., "Intelink: Using World-Wide Web Technology for Integrating Distributed Databases", 1996 IEEE Dual-Use Technology & Applications Conference, ON Center, Syracuse, NY, June 1996.

**Intersubband Heavy-Hole Scattering by Confined Optical Phonons in
Si/ZnS Superlattices:
a Design Step Towards Si/ZnS Near-Infrared Intersubband Lasers**

Gang Sun
Assistant Professor
Engineering Program/Physics Department

University of Massachusetts at Boston
100 Morrissey Blvd.
Boston, MA 02125

Final Report for:
Summer Faculty Research Extension Program
Rome Laboratory
Hanscom Air Force Base

Sponsored by:
Air Force Office of Scientific Research
Bolling Air Force Base, DC

and

Rome Laboratory

March, 1998

**Intersubband Heavy-Hole Scattering by Confined Optical Phonons in
Si/ZnS Superlattices:
a Design Step Towards Si/ZnS Near-Infrared Intersubband Lasers**

Gang Sun
Assistant Professor
Engineering Program/Physics Department
University of Massachusetts at Boston

ABSTRACT

The confinement of optical modes of vibration in a superlattice consisting of polar and nonpolar materials is described by a continuum model. Specifically, the structure under investigation is the Si/ZnS superlattice. Optical phonon modes in Si and ZnS layers are totally confined within their respective layers since both layers can be treated as infinitely rigid with respect to the other layer. Since there are no associated electric fields with nonpolar optical phonons in Si layers, only a mechanical boundary condition needs to be satisfied for these nonpolar optical modes at the Si-ZnS interface. The optical phonons in Si layers can be described by guided modes consisting of an uncoupled s-TO mode and a hybrid of LO and p-TO modes with no interface modes. In ZnS layers, a continuum model hybridizing the LO, TO and IP modes is necessary to satisfy both the mechanical and electrostatic boundary conditions at the heterointerface. A numerical procedure is provided to determine the common frequency between LO, TO, and IP modes. This is a new procedure for obtaining the eigen modes of a mixed polar-nonpolar heterosystem. Analytical expressions are obtained for the ionic displacement and associated electric field as well as scalar and vector potentials. The established model for the confined optical phonons is used in calculating the intersubband heavy-hole scattering rate by optical phonons in the Si/ZnS superlattice. Our results indicate that contributions to the intersubband scattering rate from Si or ZnS confined optical phonons depend strongly on the distribution of envelope wavefunctions over the respective layers within which different types of optical phonons are confined.

Intersubband Heavy-Hole Scattering by Confined Optical Phonons in Si/ZnS Superlattices: a Design Step Towards Si/ZnS Near-Infrared Intersubband Lasers

Gang Sun

I. INTRODUCTION

The demonstration of the InGaAs/AlInAs intersubband quantum cascade laser at $\lambda = 4.2\mu\text{m}$ [1, 2] has spurred interest in the use of silicon as the lasing material because of its integrability with advanced silicon microelectronics[3, 4]. There is also interest in moving the lasing from the far- and mid-infrared range to the near-infrared optical communications wavelengths, $\lambda = 1.3$ or $1.55\mu\text{m}$ [5]. Since the latter wavelength corresponds to a photon energy of 800meV , the $\text{Si}_{1-x}\text{Ge}_x/\text{Si}$ heterosystem is inadequate because a maximum practical valence band offset of only $\sim 500\text{meV}$ can be obtained for pseudomorphic $\text{Si}_{1-x}\text{Ge}_x$ layers at $x = 0.5 \sim 0.6$. Therefore, alternative large-bandgap, nearly lattice-matched barrier materials for Si quantum wells must be sought; materials with sufficiently large band offsets with respect to silicon. Possible candidates include, ZnS, BeSeTe, CaF_2 , SiO_x , SiO_2 , the Si/ SiO_2 superlattice, and $\gamma\text{-Al}_2\text{O}_3$, among others[5]-[7].

The Si/ZnS heterosystem has received the most attention as current advances in epitaxy technology have allowed the growth of heterostructures consisting of polar and nonpolar materials[8, 9]. The lattice mismatch of cubic ZnS with respect to Si is only 0.3%. The valence band offset has been predicted theoretically[10]-[13], while recent experiments[14] show that the value is close to 1.5eV , sufficiently large to give intersubband energy differences in the desired 800meV range. MBE growth of ZnS upon Si, and Si upon ZnS have been demonstrated[9], with the use of an As monolayer to satisfy the local bonding requirements, although the effect of the monolayer on the offsets has not been determined.

The possibility of population inversion and the operation of intersubband lasers depend critically on the lifetimes of the involved subbands. The subband lifetimes in turn are determined by nonradiative phonon scattering processes. The purpose of the present paper is to study the optical phonon modes and their interaction with carriers in the Si/ZnS system since the optical phonon scattering is considered to be dominant in the phonon scattering processes. This combination of materials is new, since it consists of both a nonpolar and polar semiconductor. Previous studies in carrier scattering by confined optical phonons in heterostructures have been focused only on one type of phonons, either polar[15]-[23] or nonpolar[24]-[27]. In the current situation involving both polar and nonpolar materials, carrier scattering by both types of phonons needs to be considered. To the best of our knowledge, there has not been any reported investigation on this mixed nature of optical phonons, their confinement effect, and their interaction with carriers in a heterostructure. In this paper, we will present a theoretical study based on the macroscopic continuum model to describe the confined optical

phonon modes and will use this model to calculate the optical-phonon scattering of heavy holes in a heterostructure consisting of polar (ZnS) and nonpolar materials (Si), as we are interested in the feasibility of constructing an intersubband laser within the valence band of the Si/ZnS heterostructure. This valence intersubband laser will likely be a quantum-parallel superlattice laser[28] or a quantum cascade laser. Our latest thinking[29] is that each of the N laser periods will consist of one square Si quantum well containing two active heavy-hole subbands. Due to the non-parabolicity of these subbands, we believe that a population inversion, localized in k -space, can be engineered between the subbands. In this investigation, we will consider a simple superlattice comprised of alternating layers of Si and ZnS, much like the flat-band superlattice of the quantum parallel laser (Fig.2(b) of [28]). The results of this study will provide the basis for the calculation of subband lifetimes required to determine laser gain and threshold.

As described below in greater detail, since the optical dispersions (frequency versus wavevector) of the silicon ($\hbar\omega_{Si} = 64meV$) and zinc sulphide ($\hbar\omega_{ZnS} = 43meV$) have no overlap, the optical phonons are assumed to be totally confined in both materials. In the silicon layers, a continuum model with double hybridization of the longitudinal optical (LO) and transverse optical (TO) modes is used to describe the vibration patterns of the guided modes[24]. The only boundary condition that needs to be satisfied in the Si layers is the vanishing of the displacements at the Si-ZnS interface, since the ZnS layers can be considered as infinitely rigid with respect to the vibrations of the Si layer. Hence, there is no interface mode in the Si layers. The situation on the ZnS layers is more complex. Following the work by Ridley[16, 17], here a continuum model is employed with hybridization of the optical LO, TO, and interface polariton (IP) modes needed to satisfy both the mechanical and electrostatic boundary conditions at the interfaces. Specifically, the electrostatic boundary conditions are the continuity of E_x , the electric field parallel to the interface, and the continuity of D_z , the displacement field normal to the interface. The mechanical boundary condition is again the vanishing of the optical displacements since the Si layers can be considered as infinitely rigid with respect to the vibrations of the ZnS layers.

Our current work provides a complete set of analytical expressions for the optical phonon dispersion relations, optical displacements, and associated scalar and vector potentials. These expressions are subsequently used in calculating the interaction of heavy holes with the confined optical phonons. In Section II, we establish a continuum model for the optical displacement modes in Si and ZnS layers satisfying both mechanical and electrical boundary conditions. In section III, we outline a numerical procedure for determining the frequency of a ZnS optical mode inducing the intersubband scattering. In Section IV, we describe the scalar and vector potentials associated with the ZnS ionic displacement modes. In Section V, we calculate the intersubband scattering rate due to the emission of Si and ZnS optical phonons since the emission process is rather significant compared to the scattering process of optical phonon absorption. In Section VI, we summarize and discuss our results and conclusions.

II. Mode Patterns

A continuum model for the optical modes in the Si/ZnS superlattice is employed. Both mechanical and electrical boundary conditions are satisfied at the heterointerfaces. Since the optical dispersion relations (frequency versus phonon wavevector) in the two bulk materials have no overlap, the phonons are taken to be confined in their respective materials. For the Si layers, the continuum model for optical phonons in nonpolar materials[24, 26] is used. Here double hybridization of the LO (longitudinal optical) and TO (transverse optical) modes is used to give the vibration patterns of the guided modes. Since the ZnS layers are infinitely rigid with respect to the vibrations of the Si layers, only the mechanical boundary condition, the vanishing of the displacements at the interfaces, has to be satisfied.

For the polar ZnS layers, the situation is more complex and an alternate continuum model[16, 17] consisting of an intermixing of confined LO, TO, and IP (interface polariton) modes is needed in order to satisfy both the electrostatic and mechanical boundary conditions. The boundary conditions which must be satisfied are (1) the continuity of E_x , the component of electric field parallel to the interface, (2) the continuity of D_z , the component of the displacement vector normal to the interface, and (3) the vanishing of the vector displacement u at the interface.

A. Modes in Si Layers

As discussed above, since the ZnS layers can be treated as infinitely rigid, the boundary condition to be satisfied in the Si layers is the vanishing of the ionic displacement of all confined vibration modes. This is an assumption of strict confinement yielding only the guided modes. As pointed out in the continuum theory[24], the ionic displacement of confined vibrations has two components: one is the hybrid of the LO and p-polarized TO (p-TO) modes, and other is the uncoupled s-polarized TO (s-TO) mode. These modes are defined as follows: If we consider a (x, z) plane containing the normal to the layers and the phonon wavevector \mathbf{Q} , then

$$\mathbf{Q} = q_x \hat{e}_x + q_z \hat{e}_z \quad (1)$$

where \hat{e}_x and \hat{e}_z are unit vectors. The p-TO mode has its displacements normal to \mathbf{Q} and in the plane, while the s-TO displacements are normal to \mathbf{Q} and perpendicular to the plane ($||\hat{e}_y$). The form of the ionic displacement, scalar, and vector potentials in one superlattice period differs from that in a neighboring period only by a phase factor proportional to the Bloch superlattice wavevector q_{SL} . Their expressions given below are obtained by taking $q_{SL} = 0$. A description of the s-TO mode is

$$u_y = e^{iq_x x} (A_{s-TO} e^{iq_z z} + B_{s-TO} e^{-iq_z z}), \quad (2)$$

while the hybrid of the LO and p-TO modes is given by

$$\begin{aligned} u_x &= e^{iq_x x} [q_x (A_{LO} e^{iq_L z} + B_{LO} e^{-iq_L z}) + q_T (A_{p-TO} e^{iq_T z} + B_{p-TO} e^{-iq_T z})], \\ u_z &= e^{iq_x x} [q_L (A_{LO} e^{iq_L z} - B_{LO} e^{-iq_L z}) - q_x (A_{p-TO} e^{iq_T z} - B_{p-TO} e^{-iq_T z})], \end{aligned} \quad (3)$$

which are confined within the Si layer with a width of d_{Si} , $0 < z < d_{Si}$. The z -components of the LO and TO wavevector have been distinguished by q_L and q_T , respectively.

Since the LO and TO modes must have the same frequency to be effectively coupled, we must satisfy the condition

$$\omega^2 = \omega_O^2 - \alpha_L^2(q_x^2 + q_L^2) = \omega_O^2 - \alpha_T^2(q_x^2 + q_T^2), \quad (4)$$

where ω_O is the bulk Si optical phonon frequency at Γ point, α_L and α_T are the sound velocities of LO and TO dispersions in Si, respectively.

Using the boundary condition that $\mathbf{u} = 0$ at the interfaces gives for the s-TO mode

$$u_y = A e^{iq_x x} \sin(q_z z), \quad \text{with } q_z = \frac{n\pi}{d_{Si}} \quad (5)$$

where $n = 1, 2, \dots$ and A is a mode coefficient. This mode does not mix with other modes.

The hybrid LO and p-TO modes admit two classes of solutions. The 'sine' solution is

$$\begin{aligned} u_x &= 2B e^{iq_x x} q_x [\cos(q_L z) - \cos(q_T z)], \\ u_z &= 2iB e^{iq_x x} [q_L \sin(q_L z) + \frac{q_x^2}{q_T} \sin(q_T z)], \end{aligned} \quad (6)$$

and the 'cosine' solution is

$$\begin{aligned} u_x &= 2iB e^{iq_x x} [q_x \sin(q_L z) + \frac{q_L q_T}{q_x} \sin(q_T z)], \\ u_z &= 2B e^{iq_x x} q_L [\cos(q_L z) - \cos(q_T z)] \end{aligned} \quad (7)$$

where

$$q_L = \frac{n_L \pi}{d_{Si}} \quad \text{and} \quad q_T = \frac{n_T \pi}{d_{Si}}, \quad (8)$$

where $n_L = 1, 2, \dots$, $n_T = 3, 4, \dots$, $n_T - n_L = 2, 4, 6, \dots$, and B is a mode coefficient. No interface modes exist in the Si layer because of the boundary condition $\mathbf{u} = 0$.

The lowest s-TO mode pattern in Eq.(5) for $q_z = \pi/d_{Si}$ is shown in Fig.1(a) within a Si layer of $d_{Si} = 40\text{\AA}$, while the hybrid patterns of the lowest p-TO and LO modes with $q_L = \pi/d_{Si}$ and $q_T = 3\pi/d_{Si}$ are shown in Figs.1(b) and 1(c) for the 'sine' and 'cosine' solutions given in Eqs.(6) and (7), respectively within the same Si layer. The strict confinement which requires the vanishing of ionic displacements at the boundaries of Si layers is clearly demonstrated for both vibration modes.

B. Modes in ZnS Layers

The boundary conditions are the continuity of E_x , D_z , and the vanishing of \mathbf{u} at the interfaces. These conditions can be satisfied by a unique linear combination of LO,

TO, and IP modes with common frequency and common in-plane wavevector q_x ,

$$\mathbf{u} = \mathbf{u}_{LO} + \mathbf{u}_{TO} + \mathbf{u}_{IP}. \quad (9)$$

We will use this hybrid expression to calculate the electrical interaction with carriers which is considerably stronger than the optical deformation potential interaction. We need consider only the displacements u_x and u_z , since u_y associated with the s-TO mode has no related electric field and therefore does not interact with carriers electrically. Once again, the expressions are obtained by taking the Bloch superlattice wavevector $q_{SL} = 0$.

For the LO mode, the ionic displacements are

$$\begin{aligned} u_x &= e^{i(q_x x - \omega t)} q_x (A_L e^{iq_L z} + B_L e^{-iq_L z}), \\ u_z &= e^{i(q_x x - \omega t)} q_L (A_L e^{iq_L z} - B_L e^{-iq_L z}). \end{aligned} \quad (10)$$

which is confined within the ZnS layer with a width of d_{ZnS} , $-d_{ZnS}/2 < z < d_{ZnS}/2$

The associated electric fields are

$$E_x = -\rho_o u_x, \quad E_z = -\rho_o u_z, \quad (11)$$

where

$$\rho_o = \frac{e^*}{\epsilon_o \Omega}, \quad (12)$$

with the effective ionic charge

$$e^{*2} = M \Omega \omega_{LO}^2 \epsilon_o^2 \left(\frac{1}{\epsilon_\infty} - \frac{1}{\epsilon_s} \right), \quad (13)$$

where M is the reduced mass, ϵ_o is the permittivity of free space, ϵ_∞ , ϵ_s are the high-frequency and static permittivities, and Ω is the volume of primitive unit cell. The scalar potential ϕ associated with the electric field $\mathbf{E} = -\nabla\phi$ is in turn given as

$$\phi = -i\rho_o e^{i(q_x x - \omega t)} (A_L e^{iq_L z} + B_L e^{-iq_L z}). \quad (14)$$

For the TO mode

$$\begin{aligned} u_x &= e^{i(q_x x - \omega t)} q_T (A_T e^{iq_T z} + B_T e^{-iq_T z}), \\ u_z &= -e^{i(q_x x - \omega t)} q_x (A_T e^{iq_T z} - B_T e^{-iq_T z}). \end{aligned} \quad (15)$$

The electric fields associated with this mode are negligible.

For the IP mode

$$\begin{aligned} u_x &= e^{i(q_x x - \omega t)} q_p (A_P e^{iq_p z} + B_P e^{-iq_p z}), \\ u_z &= i e^{i(q_x x - \omega t)} q_p (A_P e^{iq_p z} - B_P e^{-iq_p z}) \end{aligned} \quad (16)$$

The associated electric fields are

$$E_x = -\rho_p u_x, \quad E_z = -\rho_p u_z, \quad (17)$$

where

$$\rho_p = \rho_o \frac{\omega^2 - \omega_{TO}^2}{\omega_{LO}^2 - \omega_{TO}^2}, \quad (18)$$

and ω_{LO} and ω_{TO} are bulk ZnS LO and TO optical phonon frequencies at Γ point, respectively. The electric fields associated with the interface modes propagate into the Si layers although they are treated as infinitely rigid and do not contain ZnS ionic displacement.

Being a transverse electromagnetic wave, there is a vector potential \mathbf{A} associated with the electric field $\mathbf{E} = -\partial\mathbf{A}/\partial t$. Within the ZnS layers,

$$\begin{aligned} A_x &= i \frac{\rho_p}{\omega} e^{i(q_x x - \omega t)} q_p (A_P e^{iq_p z} + B_P e^{-iq_p z}), \\ A_z &= -\frac{\rho_p}{\omega} e^{i(q_x x - \omega t)} q_p (A_P e^{iq_p z} - B_P e^{-iq_p z}) \end{aligned} \quad (19)$$

While in the Si layers, a similar expression can be obtained with another set of mode coefficients, A_{p1} and B_{p1} .

Since large in-plane wavevectors are likely to be involved when dealing with carrier transitions due to optical phonons between two subbands separated with a relatively large energy, we have endeavored to obtain analytical dispersions of the LO and TO optical branches by curve-fitting the experimental bulk phonon dispersions for the entire Brillouin zone. The requirement for common frequency yields,

$$\begin{aligned} \omega &= \omega_{LO} - \beta_L (q_x^2 + q_L^2) \\ &= \omega_{TO} - \beta_T (q_x^2 + q_T^2) \\ &= \frac{c(q_x^2 + q_p^2)^{1/2}}{n(\omega)} \end{aligned} \quad (20)$$

where $\beta_L = 0.808 THz \cdot \text{\AA}^2$ and $\beta_T = 2.194 THz \cdot \text{\AA}^2$ are obtained from curve-fitting the bulk ZnS optical phonon dispersions, c is the velocity of light in vacuum, and $n(\omega)$ is the index of refraction. In the above expressions, the frequency in the ZnS layers lies between the ZnS LO and TO zone center frequencies. Since $\omega_{TO} < \omega_{LO}$, in order for the TO frequency to be equal to a LO frequency q_T must be imaginary $q_T = iq_o$, corresponding to a TO interface mode. The modes which interact most strongly with carriers are those with frequencies near the LO branch. For these modes, the value of q_o is large, and we can take the approximation

$$\tanh(q_o d_{ZnS}) \approx 1. \quad (21)$$

In the unretarded limit ($c \rightarrow \infty$), $q_x^2 + q_p^2 \approx 0$ for the IP mode. Hence, $q_p \approx iq_x$.

Applying, at the two interfaces between layers Si and ZnS in a period of the superlattice, the conditions that u_x and u_z equal to zero along with the continuity of E_x and D_z , leads to eight simultaneous equations involving the eight unknown mode coefficients ($A_L, B_L; A_T, B_T; A_P, B_P$; and A_{P1}, B_{P1}). The following two ionic displacement

mode patterns emerge for the hybrid in Eq.(9) taking the Bloch superlattice wavevector $q_{SL} = 0$ and the approximation $\tanh(q_o d_{ZnS}) \approx 1$. Both ionic displacement patterns are confined within the ZnS layer, $-d_{ZnS}/2 < z < d_{ZnS}/2$. For the first type,

$$\begin{aligned} u_x &= 2iBe^{iq_x x} q_x \left\{ \sin(q_L z) \right. \\ &\quad - [1 - p_1 \tanh(q_x d_{ZnS}/2)] \sin(q_L d_{ZnS}/2) \frac{\sinh(q_o z)}{\sinh(q_o d_{ZnS}/2)} \\ &\quad \left. - p_1 \sin(q_L d_{ZnS}/2) \frac{\sinh(q_x z)}{\cosh(q_x d_{ZnS}/2)} \right\}, \\ u_z &= 2Be^{iq_x x} q_L \left\{ \cos(q_L z) \right. \\ &\quad - \frac{q_x^2}{q_L q_o} [1 - p_1 \tanh(q_x d_{ZnS}/2)] \sin(q_L d_{ZnS}/2) \frac{\cosh(q_o z)}{\sinh(q_o d_{ZnS}/2)} \\ &\quad \left. - \frac{q_x}{q_L} p_1 \sin(q_L d_{ZnS}/2) \frac{\cosh(q_x z)}{\cosh(q_x d_{ZnS}/2)} \right\}, \end{aligned} \quad (22)$$

and for the second type,

$$\begin{aligned} u_x &= 2Be^{iq_x x} q_x \left\{ \cos(q_L z) \right. \\ &\quad - [1 - p_2 \coth(q_x d_{ZnS}/2)] \cos(q_L d_{ZnS}/2) \frac{\cosh(q_o z)}{\sinh(q_o d_{ZnS}/2)} \\ &\quad \left. - p_2 \cos(q_L d_{ZnS}/2) \frac{\cosh(q_x z)}{\sinh(q_x d_{ZnS}/2)} \right\}, \\ u_z &= 2iBe^{iq_x x} q_L \left\{ \sin(q_L z) \right. \\ &\quad + \frac{q_x^2}{q_L q_o} [1 - p_2 \coth(q_x d_{ZnS}/2)] \cos(q_L d_{ZnS}/2) \frac{\sinh(q_o z)}{\sinh(q_o d_{ZnS}/2)} \\ &\quad \left. + \frac{q_x}{q_L} p_2 \cos(q_L d_{ZnS}/2) \frac{\sinh(q_x z)}{\sinh(q_x d_{ZnS}/2)} \right\}, \end{aligned} \quad (23)$$

where

$$\begin{aligned} p_1 &= \frac{\cosh(q_x d_{Si}/2) \cosh(q_x d_{ZnS}/2)}{d}, \\ p_2 &= \frac{\sinh(q_x d_{Si}/2) \sinh(q_x d_{ZnS}/2)}{d}, \\ d &= r \sinh(q_x d_{Si}/2) \cosh(q_x d_{ZnS}/2) + \sinh(q_x d_{ZnS}/2) \cosh(q_x d_{Si}/2), \\ r &= \frac{\epsilon_{p1}}{\epsilon_{p2}}, \end{aligned} \quad (24)$$

and ϵ_{p1} and ϵ_{p2} are the permittivities in Si and ZnS layers, respectively, with

$$\epsilon_{p2} = \epsilon_\infty \frac{\omega^2 - \omega_{LO}^2}{\omega^2 - \omega_{TO}^2}. \quad (25)$$

To illustrate the patterns of ionic displacements in the ZnS layers given in Eqs.(22) and (23), we need to first determine values for q_x , q_L , and q_o . To do so, we will follow the numerical procedure described in Section III by arbitrarily fixing a value for the in-plane phonon wavevector $q_x = \pi/(5a_{ZnS})$, where a_{ZnS} is the lattice constant of ZnS. In

calculating the carrier-optical phonon interaction, the value of q_x is actually determined by the conservation of in-plane momentum between the initial and final states of the scattering process. For a given value of q_x , typically, a set of hybridized modes can be obtained. Here, we show only the mode pattern with frequency close to ω_{LO} .

We obtained $\hbar\omega = 41\text{meV}$, $q_L = 0.46 \times 10^8/\text{cm}$ and $q_o = 0.48 \times 10^8/\text{cm}$. Substituting these values into Eqs.(22) and (23), we obtained Figs.2(a) and 2(b) showing the mode patterns of ionic displacement of both the first and second types, respectively in a ZnS layer of $d_{ZnS} = 20\text{\AA}$. It can be seen from Figs.2(a) and 2(b) that the mechanical boundary condition, vanishing of the ionic displacements at the interfaces of Si and ZnS layers, is satisfied.

III. Dispersion Relationship

The phonon frequency in the ZnS layers is determined by the following set of equations;

$$\begin{cases} \omega = \omega_{LO} - \beta_L(q_x^2 + q_L^2), \\ \omega = \omega_{TO} - \beta_T(q_x^2 - q_o^2), \\ t_1 + t_2 \cos(q_L d_{ZnS}) + t_3 \sin(q_L d_{ZnS}) = 0 \end{cases} \quad (26)$$

where

$$\begin{aligned} t_1 &= 4p \sinh(q_x d_{Si}) + 4pr \sinh(q_x d_{ZnS}), \\ t_2 &= -4p\alpha, \\ t_3 &= 8p^2 r \sinh(q_x d_{ZnS}) \sinh(q_x d_{Si}) - 4p^2 \alpha^2 \\ &\quad + 4p^2 r^2 \sinh^2(q_x d_{ZnS}) + 4p^2 \sinh^2(q_x d_{Si}) + 1, \end{aligned} \quad (27)$$

with

$$p = \frac{q_x}{4q_L r s d} \quad (28)$$

and

$$s = \frac{\omega^2 - \omega_{TO}^2}{\omega_{LO}^2 - \omega_{TO}^2}. \quad (29)$$

The third equation in (26) is obtained from the requirement of a nonzero solution for the eight simultaneous equations discussed above, and Eq.(27) is arrived at under the approximation, $\tanh(q_o d_{ZnS}) \approx 1$.

The numerical procedure for determining a phonon frequency is the following: given a value of q_x , we can determine those of t_1 , t_2 , and t_3 from Eq.(27). Then ω is scanned from ω_{TO} to ω_{LO} . For a given value of ω , q_L and q_o are obtained from the first two equations in (26). Those values are then substituted into the third equation in (26) to determine if the particular value of ω is a solution.

IV. Scalar and Vector Potentials

Associated with the two types of ionic displacement in Eqs.(22) and (23), the scalar potentials in ZnS layers are given as, for the first type,

$$\phi = \begin{cases} 2\rho_o B e^{iq_x x} \sin(q_L z_1) & |z_1| < \frac{d_{ZnS}}{2} \quad \text{ZnS layer,} \\ 0 & |z_2| < \frac{d_{Si}}{2} \quad \text{Si layer,} \end{cases} \quad (30)$$

and for the second type,

$$\phi = \begin{cases} -2i\rho_o B e^{iq_x x} \cos(q_L z_1) & |z_1| < \frac{d_{ZnS}}{2} \quad \text{ZnS layer,} \\ 0 & |z_2| < \frac{d_{Si}}{2} \quad \text{Si layer.} \end{cases} \quad (31)$$

Note that we have used two different coordinates z_1 and z_2 for layers ZnS and Si, respectively, with their origins placed at the centers of the respective layers.

The vector potentials can be obtained, for the first type,

$$A_x = \begin{cases} \frac{2s\rho_o q_x}{\omega} B e^{iq_x x} p_1 \sin(q_L d_{ZnS}/2) \frac{\sinh(q_x z_1)}{\cosh(q_x d_{ZnS}/2)} & |z_1| < \frac{d_{ZnS}}{2} \quad \text{ZnS layer,} \\ \frac{4q_x \rho_o}{\omega} B e^{iq_x x} V_1 \sinh(q_x z_2) & |z_2| < \frac{d_{Si}}{2} \quad \text{Si layer,} \end{cases} \quad (32)$$

$$A_z = \begin{cases} \frac{2is\rho_o q_x}{\omega} B e^{iq_x x} p_1 \sin(q_L d_{ZnS}/2) \frac{\cosh(q_x z_1)}{\cosh(q_x d_{ZnS}/2)} & |z_1| < \frac{d_{ZnS}}{2} \quad \text{ZnS layer,} \\ \frac{4iq_x \rho_o}{\omega} B e^{iq_x x} V_1 \cosh(q_x z_2) & |z_2| < \frac{d_{Si}}{2} \quad \text{Si layer,} \end{cases} \quad (33)$$

and for the second type,

$$A_x = \begin{cases} -\frac{2is\rho_o q_x}{\omega} B e^{iq_x x} p_2 \cos(q_L d_{ZnS}/2) \frac{\cosh(q_x z_1)}{\sinh(q_x d_{ZnS}/2)} & |z_1| < \frac{d_{ZnS}}{2} \quad \text{ZnS layer,} \\ \frac{4iq_x \rho_o}{\omega} B e^{iq_x x} V_2 \cosh(q_x z_2) & |z_2| < \frac{d_{Si}}{2} \quad \text{Si layer,} \end{cases} \quad (34)$$

$$A_z = \begin{cases} -\frac{2s\rho_o q_x}{\omega} B e^{iq_x x} p_2 \cos(q_L d_{ZnS}/2) \frac{\sinh(q_x z_1)}{\sinh(q_x d_{ZnS}/2)} & |z_1| < \frac{d_{ZnS}}{2} \quad \text{ZnS layer,} \\ \frac{4q_x \rho_o}{\omega} B e^{iq_x x} V_2 \sinh(q_x z_2) & |z_2| < \frac{d_{Si}}{2} \quad \text{Si layer,} \end{cases} \quad (35)$$

where

$$\begin{aligned} V_1 &= \frac{r \sin(q_L d_{ZnS}/2) \cosh(q_x d_{ZnS}/2)}{2d}, \\ V_2 &= \frac{r \cos(q_L d_{ZnS}/2) \sinh(q_x d_{ZnS}/2)}{2d}, \end{aligned} \quad (36)$$

and p_1 , p_2 , and d are given in Eq.(24).

The scalar potentials associated with the LO modes are strictly confined within the ZnS layers. Their distributions are shown in Fig.3 for the first and second types given in Eqs.(30) and (31) with $q_L = 0.31 \times 10^8/cm$, $d_{ZnS} = 20\text{\AA}$, respectively.

The vector potential associated with the IP modes are distributed in both Si and ZnS layers, even though Si layers are treated as infinitely rigid and do not contain ZnS ionic displacements. The profiles for the two components of the vector potentials given in Eqs.(32-35) for the first and second types with $d_{Si} = 40\text{\AA}$, $d_{ZnS} = 20\text{\AA}$ are shown in Figs.4(a) and 4(b), respectively.

It can be seen from Figs.3 and 4 that both scalar and vector potentials are not continuous across the interfaces. However, as pointed by Ridley[16], the energy of interaction with an electron traveling coherently with the optical phonon is continuous. The electric field can be obtained as

$$\mathbf{E} = -\nabla\phi - \frac{\partial \mathbf{A}}{\partial t}. \quad (37)$$

The continuity of E_x and $D_z = \epsilon(\omega)E_z$ implies that at the boundaries,

$$\begin{aligned} \omega A_x|_{z_2=\pm d_{Si}/2} &= -q_x \phi|_{z_1=\mp d_{ZnS}/2} + \omega A_x|_{z_1=\mp d_{ZnS}/2}, \\ A_z|_{z_2=\pm d_{Si}/2} &= r A_z|_{z_1=\mp d_{ZnS}/2}, \end{aligned} \quad (38)$$

where A_{1x} and A_{1z} are x - and z -components of the vector potential in Si layers. The interaction in the Si layer is $e(A_{1x}v_x + A_{1z}v_z)$ and in the ZnS layer $e(-\phi + A_xv_x + A_zv_z)$, which are equal when the electron velocity $v_x = \omega/q_x$ and $v_z = 0$. Thus, the coherent interaction energy is continuous across the interfaces.

The electric field distributions for E_x and $\epsilon(\omega)$ in Si ($d_{Si} = 40\text{\AA}$) and ZnS ($d_{ZnS} = 40\text{\AA}$) layers are shown in Figs.5(a) and 5(b) for the first and second types, respectively. The continuity of E_x and D_z across the Si and ZnS interface according to Eq.(38) is clearly demonstrated.

V. Intersubband Scattering

Since the optical modes in the Si/ZnS superlattice consist of confined nonpolar Si and polar ZnS optical phonons, the calculation of carrier scattering by optical phonons in such a structure needs to include contributions from both types of phonons. The Hamiltonian that describes the carrier interaction with the nonpolar Si optical phonons is given by[30]

$$H = \frac{1}{2} \mathbf{D} \cdot \mathbf{u} \quad (39)$$

where \mathbf{D} is the optical deformation potential. This Hamiltonian obviously vanishes outside of the Si layers in the Si/ZnS superlattice since the Si optical displacement modes are strictly confined within the Si layers. However, the carrier interaction with the confined polar ZnS optical phonons extends over both ZnS and Si layers. The

electrical interaction Hamiltonian can be obtained using the scalar and vector potentials

$$H = -e\phi + \frac{e}{m} \mathbf{A} \cdot \mathbf{p}, \quad (40)$$

where \mathbf{p} is the momentum operator, e and m are the free electron charge and mass, respectively. Although the scalar potential ϕ associated with the LO mode vanishes in Si layers, the $\mathbf{A} \cdot \mathbf{p}$ interaction exists in both layers since the vector potential associated with the interface modes in the ZnS layers as shown in Fig.4 propagates into the Si layers as well.

A. Scattering due to Si Phonons

The displacement patterns described in Eqs.(5-7) all contain an arbitrary constant for the mode amplitude which can be normalized by equating the energy of the vibration mode with that of a simple harmonic oscillator[16]

$$\chi^2 = \frac{S}{\Omega} \int_0^{d_{Si}} \mathbf{u}^* \cdot \mathbf{u} dz, \quad (41)$$

where S is the sample surface area (in (x, y) plane), Ω is the volume of the unit cell, and χ is the normal coordinator of the oscillator. The heavy-hole state can be characterized by $|\mathbf{k}, n\rangle$ with the in-plane momentum \mathbf{k} and subband index n . In the approximation of constant effective mass for heavy holes, the matrix element for the transition from state $|\mathbf{k}, n\rangle$ to $|\mathbf{k}', n'\rangle$ due to the emission of a nonpolar Si optical phonon is

$$\langle \mathbf{k}', n' | H | \mathbf{k}, n \rangle = \begin{cases} \sqrt{\frac{\hbar[n(\omega_o) + 1]}{2\rho_{Si}\omega_o S d_{Si} \Delta_A(q_z)}} \delta_{\mathbf{k}' \pm \mathbf{q}_x, \mathbf{k}} D_y G_{nn'}^y(q_z) & \text{(s-TO)} \\ \sqrt{\frac{\hbar[n(\omega_o) + 1]}{2\rho_{Si}\omega_o S d_{Si} \Delta_C(q_L, q_T)}} \delta_{\mathbf{k}' \pm \mathbf{q}_x, \mathbf{k}} \\ \cdot [D_x G_{nn'}^x(q_L, q_T) + D_z G_{nn'}^z(q_L, q_T)] & \text{(hybrid)}, \end{cases} \quad (42)$$

for the s-TO mode and the hybrid of the LO and p-TO mode, respectively. $n(\omega_o)$ is the number of Si optical phonons at thermal equilibrium, and ρ_{Si} is the density of Si. The three components of the optical deformation potential, D_x , D_y , and D_z are assumed equal to $D_o = D/\sqrt{3}$ in the calculation, in view of the assumption of isotropy. The Kronecker symbol indicates the in-plane (x, y) momentum conservation. The normalization factors are given by

$$\begin{aligned} \Delta_A(q_z) &= \frac{1}{d_{Si}} \int_0^{d_{Si}} u_y^* u_y dz & \text{(s-TO),} \\ \Delta_C(q_L, q_T) &= \frac{1}{d_{Si}} \int_0^{d_{Si}} (u_x^* u_x + u_z^* u_z) dz & \text{(hybrid).} \end{aligned} \quad (43)$$

The $G_{nn'}$ -functions contain envelope wavefunctions, ψ_n and ψ'_n , from which interference effect can be obtained. Specifically,

$$G_{nn'}^y(q_z) = \int_0^{d_{Si}} \psi_n^* \psi_{n'} u_y dz, \quad (44)$$

for the s-TO mode, and

$$\begin{aligned} G_{nn'}^x(q_L, q_T) &= \int_0^{d_{Si}} \psi_n^* \psi_{n'} u_x dz, \\ G_{nn'}^z(q_L, q_T) &= \int_0^{d_{Si}} \psi_n^* \psi_{n'} u_z dz, \end{aligned} \quad (45)$$

for the hybrid of LO and p-TO mode. The heavy-hole energy levels and envelope wavefunctions are obtained by the finite square well model for the superlattice with the heavy-hole band offset taken to be $1.5eV$ [14].

Applying the Fermi golden rule, we obtain the scattering rate due to the emission of a nonpolar Si optical phonon,

$$W_{nn'} = \begin{cases} \frac{m_{hh}^*[n(\omega_o) + 1]D_o^2}{2\hbar^2 \rho_1 \omega_o d_{Si}} \sum_{q_z} \frac{|G_{nn'}^y|^2}{\Delta_A} & (\text{s-TO}) \\ \frac{m_{hh}^*[n(\omega_o) + 1]D_o^2}{2\hbar^2 \rho_1 \omega_o d_{Si}} \sum_{q_L, q_T} \frac{|G_{nn'}^x + G_{nn'}^z|^2}{\Delta_C} & (\text{hybrid}), \end{cases} \quad (46)$$

where we have assumed that for the intersubband process ($n \neq n'$) the heavy holes are scattered from the bottom of their original subbands ($\mathbf{k} = 0$), and the sum is over those participating modes of Eq.(8) that, according to Eq.(4), yield values of q_x satisfying the in-plane momentum conservation[24].

B. Scattering due to ZnS Phonons

The normalization of the amplitudes of the confined ZnS displacement modes can be carried out by equating the energy of a hybrid, a mixture of mechanical and electromagnetic energies, with that of a simple harmonic oscillator[16]. Since only the IP mode contributes electromagnetic energy which is small in magnitude when compared with the mechanical energy, neglecting it entirely will introduce little error in evaluating the energy of a hybrid ZnS mode. We therefore can use Eq.(41) for the normalization of a ZnS mode, except that now the integral is over the ZnS layer, d_{ZnS} . The matrix element for the transition from state $|\mathbf{k}, n\rangle$ to $|\mathbf{k}', n'\rangle$ due to the emission of a polar ZnS optical phonon is

$$\langle \mathbf{k}', n' | H | \mathbf{k}, n \rangle = \sqrt{\frac{e^2 \hbar [n(\omega_{ZnS}) + 1] \omega_{ZnS}}{2S d_{ZnS} \epsilon_P \Delta_{1,2}}} \delta_{\mathbf{k}' \pm \mathbf{q}, \mathbf{k}} \left(-G_{nn'}^\phi + \frac{\hbar k_x}{m} G_{nn'}^x - \frac{i\hbar}{m} G_{nn'}^z \right) \quad (47)$$

for both the first and second types. $n(\omega_{ZnS})$ is the number of ZnS optical phonons at thermal equilibrium, and

$$\frac{1}{\epsilon_p} = \frac{1}{\epsilon_\infty} - \frac{1}{\epsilon_s}. \quad (48)$$

The normalization factors for both the first and second types can all be calculated by

$$\Delta_{1,2} = \frac{1}{d_{ZnS}} \int_0^{d_{ZnS}} (u_x^* u_x + u_z^* u_z) dz \quad (49)$$

with optical displacements given in Eqs.(22) and (23). The $G_{nn'}$ -functions containing the interference effect between two subband envelope wavefunction, ψ_n and $\psi_{n'}$, are given specifically as

$$G_{nn'}^\phi = \frac{1}{\rho_o} \int_0^{d_{ZnS}} \phi \psi_n^* \psi_{n'} dz, \quad (50)$$

for the scalar potential scattering associated with the LO modes, and

$$\begin{aligned} G_{nn'}^x &= \frac{1}{\rho_o} \int_0^{d_{ZnS}} A_x \psi_n^* \psi_{n'} dz, \\ G_{nn'}^z &= \frac{1}{\rho_o} \int_0^{d_{ZnS}} A_z \psi_n^* \frac{\partial}{\partial z} \psi_{n'} dz, \end{aligned} \quad (51)$$

for the vector potential scattering associated with the IP modes. Applying the Fermi golden rule, the intersubband scattering rate due to the emission of a polar ZnS phonon can then be obtained by taking summation over contributing confined ZnS optical modes

$$W_{nn'} = \frac{e^2 m_{hh}^* [n(\omega_{LO}) + 1] \omega_{LO}}{2 \hbar^2 d_{ZnS} \epsilon_p} \sum_{q_L} \frac{|-G_{nn'}^\phi + \frac{\hbar k_x}{m} G_{nn'}^x - \frac{i \hbar}{m} G_{nn'}^z|^2}{\Delta_{1,2}}, \quad (52)$$

where we have again assumed that the heavy holes are scattered from the bottom of subband n ($k = 0$), and have taken the approximation of $\omega_{ZnS} = \omega_{LO}$ since the modes which interact most strongly with carriers are those with frequencies near the LO branch.

C. Intersubband Scattering Rates

The scattering rates due to the emission of Si and ZnS optical phonons were calculated for the intersubband transition (2-1) originated from the bottom of the heavy-hole subband 2 with zero kinetic energy to heavy-hole subband 1. Figure 6 shows the 2-1 scattering rates as a function of the Si well width while fixing the barrier width at $d_{ZnS} = 40 \text{ \AA}$ in the Si/ZnS superlattice. The total scattering rate is the summation of contributions from the heavy-hole interaction with Si and ZnS optical phonons. In the small well width region ($d_{Si} < 30 \text{ \AA}$), the heavy-hole scattering due to the ZnS optical phonons is stronger than that due to the Si optical phonons. This is attributed to the fact that when the Si well width is small there is significant envelope function overlap between subbands 1 and 2 in the ZnS barrier region where the ZnS LO phonons are confined. As the well width increases, the distribution of envelope functions in the barrier decreases. As a result, the scattering due to the ZnS LO phonons reduces considerably and the ZnS phonon scattering is mostly through IP modes which propagate throughout the superlattice structure. As the well width continues to increase, the energy separation between subband 1 and 2 decreases. The intersubband scattering between these two subbands requires an emitted ZnS phonon with a small in-plane wavevector (q_x) in order to satisfy the in-plane momentum conservation for the scattering process to take place. This leads to an increased intersubband scattering rate since polar optical phonons with smaller wavevectors interact more strongly with carriers to induce intersubband transitions as

suggested by the well-known $1/(q_x^2 + q_z^2)$ dependence of the interaction Hamiltonian in polar material quantum wells[15]. A similar dependence of the intersubband scattering rate of Eq.(52) due to the confined ZnS optical phonons is implicitly included in the normalization factor ($\Delta_{1,2}$) given by Eq.(49). Further increasing the well width to $d_{Si} > 82\text{\AA}$ causes the energy separation between the two subbands to be less than the ZnS optical phonon energy (43meV) and the heavy holes at the bottom of subband 2 cannot emit ZnS optical phonons to make a transition to subband 1, resulting in zero scattering rate due to the emission of ZnS optical phonon. The scattering rate due to the emission of Si optical phonon confined within the Si well demonstrates a steady decrease as the well width (d_{Si}) increases, which suggests that the factor $1/d_{Si}$ in Eq.(46) dominates the small increase in the interference effect $G_{nn'}$ function. As the well width increases beyond 62\AA , the energy separation between the two heavy-hole subbands becomes less than the Si optical phonon energy (64meV). As a result, the scattering rate due to the emission of Si optical phonons reduces to zero, in which case the heavy-hole lifetime of subband 2 can be enhanced dramatically since the significant scattering process of optical phonon emission is suppressed although the weaker optical phonon absorption and acoustic phonon scattering processes are still possible.

Figure 7 shows the intersubband scattering rates between the same two heavy-hole subbands due to the emission of both Si and ZnS optical phonons as a function of the barrier width (d_{ZnS}) in the Si/ZnS superlattice. The well width (d_{Si}) is fixed at 30\AA . The scattering rate due to the emission of Si optical phonons remains unchanged as the barrier width varies since the subband energy levels are hardly shifted and the $G_{nn'}$ -function for the Si phonon scattering has little noticeable change. The scattering rate due to the emission of an ZnS optical phonon, on the other hand, demonstrates a decreasing trend as the ZnS barrier width increases as suggested in Eq.(52) with the factor of $1/d_{ZnS}$. The small discontinuous incremental steps in the ZnS-scattering curve are due to the discrete nature of the increase in the number of allowed LO modes confined in the ZnS barrier as it increases.

VI. Summary and Discussion

We have provided an analytical model of optical modes in Si/ZnS superlattices consisting of polar and nonpolar optical phonons. This is a new procedure for obtaining the eigen modes of a mixed polar-nonpolar heterosystem. In the Si layers, a continuum model with double hybridization of the LO and TO modes is used to describe the vibration patterns. Since there is no electric field resulting from the nonpolar ionic displacements in Si layers, the only boundary condition that needs to be satisfied in the Si layers is the vanishing of the displacements at the Si-ZnS interface, as the ZnS layers can be considered as infinitely rigid with respect to the vibrations of the Si layer. Due to this strict confinement, only guided modes emerge in the Si layers which consist of s-TO and coupled p-TO and LO modes, with no interface modes. These guided modes have been illustrated. Their interaction with carriers in the superlattice can be calculated through the optical deformation potential for Si. The interaction Hamiltonian can be

obtained by taking the product of this potential with the normalized ionic displacement.

However, for the optical phonons in ZnS layers, we need to include the electrical interaction in calculating the carrier scattering by optical phonons, since there are electric fields associated with the polar optical vibrations. As a result, both mechanical and electrostatic boundary conditions need to be satisfied in the interfaces. A continuum model employing a linear combination of LO, TO and IP (interface polariton) modes with a common frequency is used to describe the ionic displacements in ZnS layers. A numerical procedure for determining a phonon frequency is provided. This hybridized model is necessary to meet the simultaneous requirement on the mechanical and electrostatic boundary conditions. The mechanical boundary condition is again the vanishing of the optical displacements since Si layers can be considered as infinitely rigid with respect to the vibrations of the ZnS layers. The electrostatic boundary conditions are the continuity of the electric field parallel to the interface, and the continuity of the displacement field normal to the interface. Based on this set of boundary conditions, expressions are obtained for the ionic displacements in ZnS layers consisting of LO, TO, and IP modes. There are scalar and vector potentials associated with the LO and IP modes, respectively, but no electric field associated with the TO mode. The scalar potential and its associated electric field due to the LO mode are distributed only within the ZnS layers and are zero in the Si layers. But the vector potential and its associated electric field due to the IP mode have distributions in both ZnS and Si layers even though there is no ZnS ionic displacement mode in the Si layers. Examples of these mode characteristics have been demonstrated. Neither the scalar nor vector potential is continuous across the Si-ZnS interface, but the energy of coherent interaction with carriers is continuous due to the continuity of the electric field parallel to the interface.

The analytical model for the confined optical modes consisting of polar and non-polar optical phonons is employed in calculating the carrier-phonon interaction. Our results indicate that contributions to heavy-hole intersubband scattering from confined Si and ZnS optical phonons strongly depend on the well width since it varies the distributions of envelope functions of involved subbands which ultimately determines the intersubband scattering between them through the overlapping interference effect $G_{nn'}$ -function. For small Si well width ($< 30\text{\AA}$), the scattering rate due to ZnS optical phonon is stronger than that of Si optical phonons. As the well width increases the scattering rate due to the Si optical phonons surpasses that of ZnS optical phonons. The scattering rate dependence on barrier width is relatively weak.

References

- [1] J. Faist, F. Capasso, D. L. Sivco, A. L. Hutchinson, C. Sirtory, and A. Y. Cho, *Science* **264**, 553 (1994)
- [2] J. Faist, F. Capasso, D. L. Sivco, A. L. Hutchinson, C. Sirtory, S. N. G. Chu, and A. Y. Cho, *Appl. Phys. Lett.* **65**, 2091 (1994)
- [3] G. Sun, L. Friedman, and R.A. Soref, *Appl. Phys. Lett.* **66**, 3425 (1995)
- [4] R. A. Soref, *Proc. IEEE* **81**, 1687 (1993)
- [5] L. Friedman and R. A. Soref, *IEEE Photonics Technology Letters* **5**, 1200 (1993)
- [6] L. J. Schowalter and R. W. Fathauer, *CRC Critical Review* **15**, 367 (1989)
- [7] R. Tsu, *Nature* **364**, 19 (1993)
- [8] M Yokoyama, K. I. Kashiro, and S. I. Ohta, *J. Crystal Growth* **81**, 73 (1987)
- [9] X. Zhou, S. Jiang, and W. P. Kirk, *J. Appl. Phys.* **82**, 2251-2262 (1997)
- [10] E. G. Wang and C. S. Ting, *Phys. Rev. B* **51**, 9791 (1995)
- [11] C. Maierhofer, S. Kulkarni, M. Alonso, T. Reich, and K. Horn, *J. Vac. Sci. Technol. B* **9**, 2238 (1991)
- [12] M. Cardona and N. E. Christensen, *J. Vac. Sci. Technol. B* **6**, 1285 (1988)
- [13] W. A. Harrison, *J. Vac. Sci. Technol.* **14**, 1016 (1977)
- [14] S. Jiang, X. Zhou, T. Zhou, K. P. Clark, G. Spencer, R. T. Bate, W. P. Kirk, R. M. Steinhoff, and B. Brar, submitted to *J. Appl. Phys.* (Sept., 1997)
- [15] B. K. Ridley, *Phys. Rev. B* **39**, 5282 (1989)
- [16] B. K. Ridley, *Phys. Rev. B* **47**, 4592 (1993)
- [17] M. P. Chamberlain, M Cardona, and B. K. Ridley, *Phys. Rev. B* **48**, 14356 (1993)
- [18] N. C. Constantinou and B. K. Ridley, *Phys. Rev. B* **49**, 17065 (1994)
- [19] B. K. Ridley, *Appl. Phys. Lett.* **66**, 3633 (1995)
- [20] E. Molinari and A. Fasolino, *Appl. Phys. Lett.* **54**, 1220 (1989)
- [21] L. register, *Phys. Rev. B* **45**, 8756 (1992)
- [22] K. J. Nash, *Phys. Rev. B* **46**, 7723 (1992)
- [23] N. Mori and T. Ando, *Phys. Rev. B* **40**, 6175 (1989)
- [24] G. Sun and L. Friedman, *Phys. Rev. B* **53**, 3966 (1995)

- [25] A. Fasolino, E. Molinari, and J. C. Mann, Phys. Rev. B **39**, 3923 (1989)
- [26] B. K. Ridley, Phys. Rev. B **44**, 9002 (1991)
- [27] S. C. Jain and W. Hayes, Semicond. Sci. Technol. **6**, 547 (1991)
- [28] L. Friedman, R. A. Soref, and G. Sun, IEEE Photonics Technology Letters, **9**, 593 (1997)
- [29] R. A. Soref, L. Friedman, L. C. Lew Yan Voon, L. R. Ram-Mohan, and G. Sun, submitted to J. Vac. Sci. Technol. B (Sept., 1997)
- [30] B. K. Ridley, Quantum Processes in Semiconductors (Clarendon Press, Oxford, 1982), Chap. 3

Figure Captions

Figure 1. Vibration patterns in a Si layer with a width of 40\AA for (a) the guided s-TO mode, (b) the 'sine' solution, and (c) the 'cosine' solution of the guided p-TO and LO modes.

Figure 2. Vibration patterns in a ZnS layer with a width of 20\AA for (a) the first type and (b) the second type solutions of the hybridized LO, TO and IP modes.

Figure 3. Scalar potential distribution associated with the LO modes in a period of the Si/ZnS superlattice with $d_{Si} = 40\text{\AA}$ and $d_{ZnS} = 20\text{\AA}$ for both the first and second types of the vibration modes.

Figure 4. Vector potentials associated with the IP modes distributed in a period of the Si/ZnS superlattice with $d_{Si} = 40\text{\AA}$ and $d_{ZnS} = 20\text{\AA}$ for (a) the first type and (b) the second type of the vibration modes.

Figure 5. The field distributions, E_x and D_z , derived from the scalar and vector potentials, in a period of the Si/ZnS superlattice with $d_{Si} = 40\text{\AA}$ and $d_{ZnS} = 20\text{\AA}$ for (a) the first type and (b) the second type of the vibration modes.

Figure 6. Intersubband scattering rates due to the emission of Si and ZnS optical phonons as a function of Si well width (d_{Si}) for a barrier width of $d_{ZnS} = 40\text{\AA}$.

Figure 7. Intersubband scattering rates due to the emission of Si and ZnS optical phonons as a function of ZnS barrier width (d_{ZnS}) for a well width of $d_{Si} = 30\text{\AA}$.

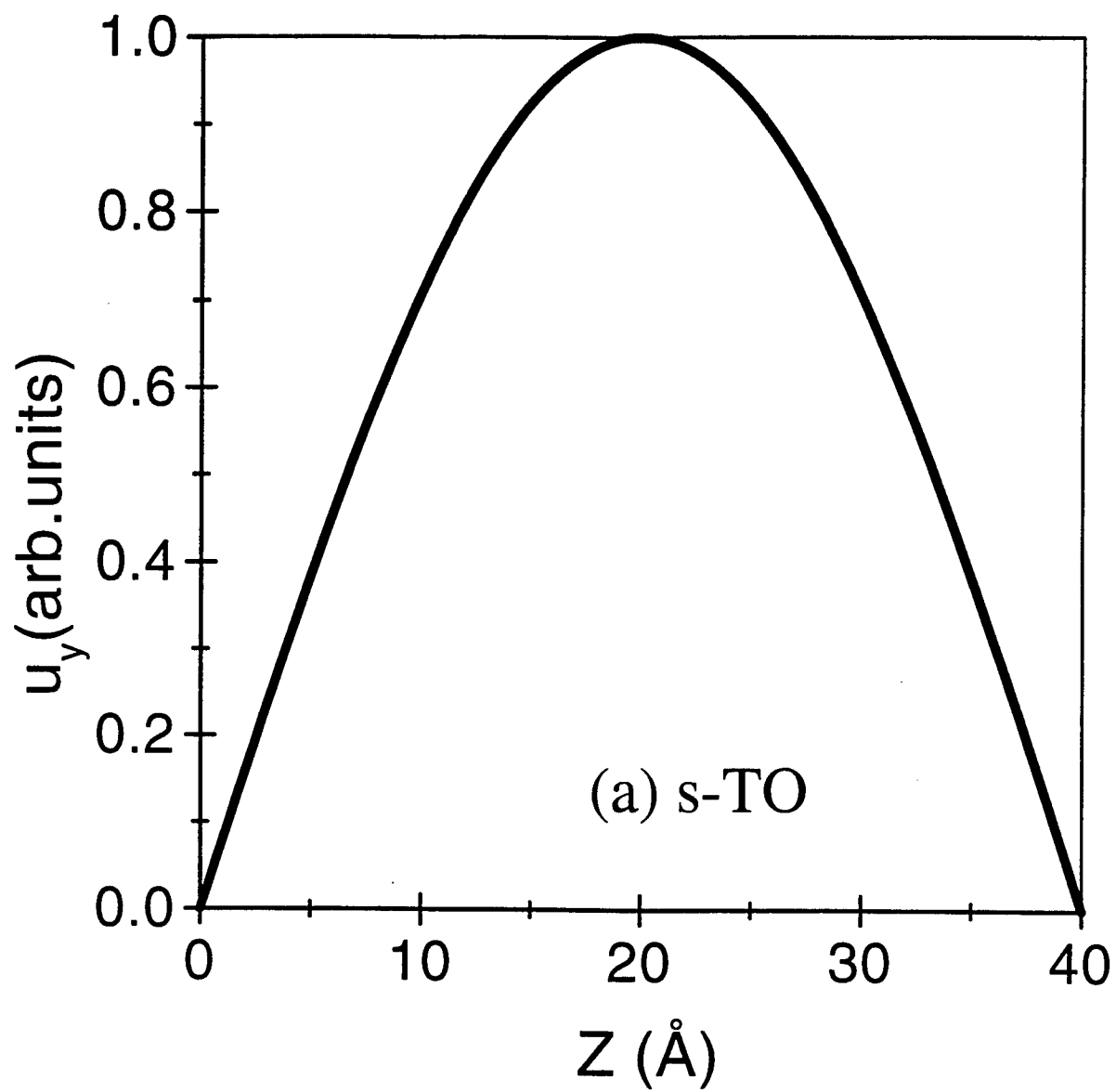


Fig. 1(a)

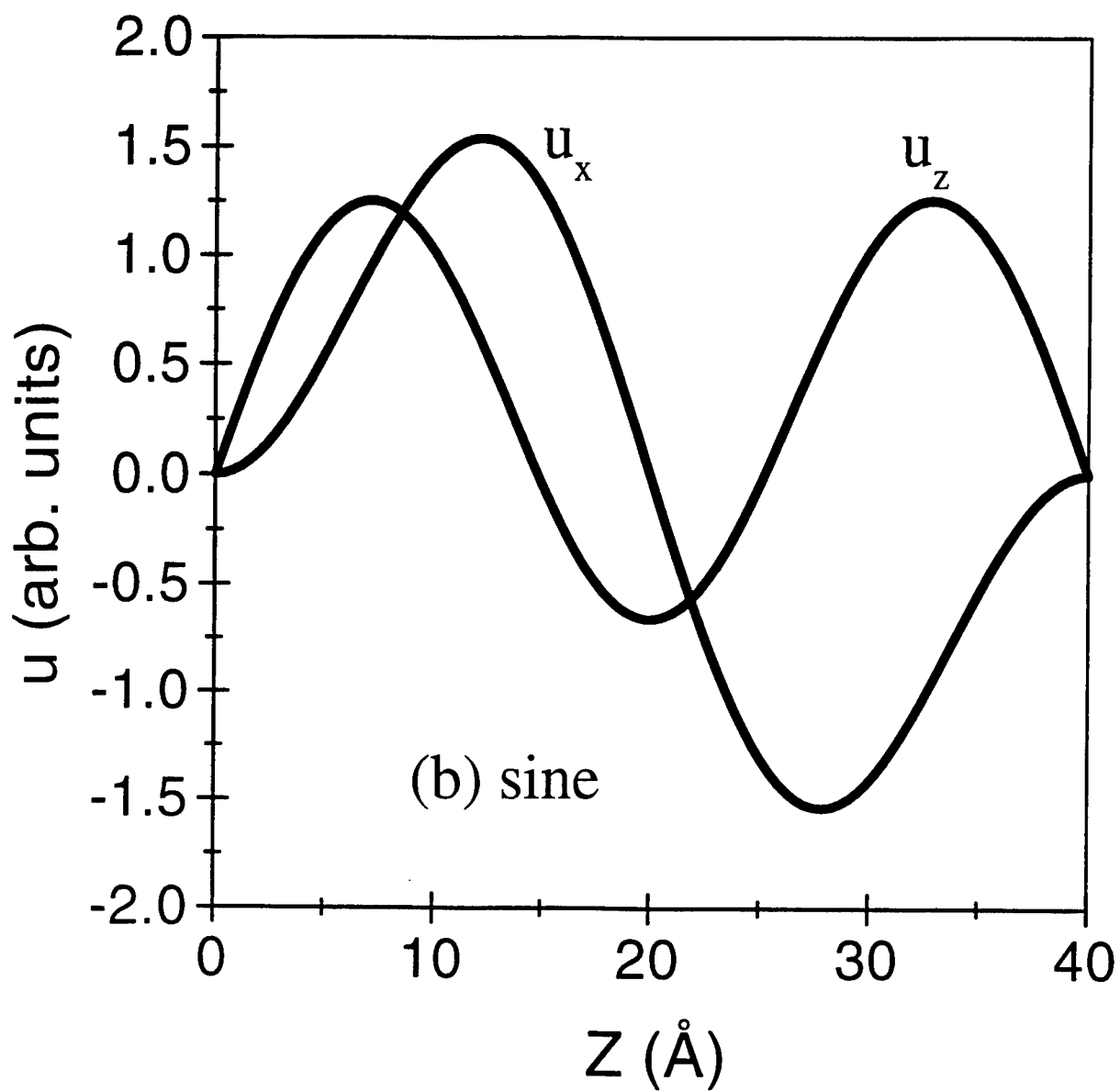


Fig. 1 (b)

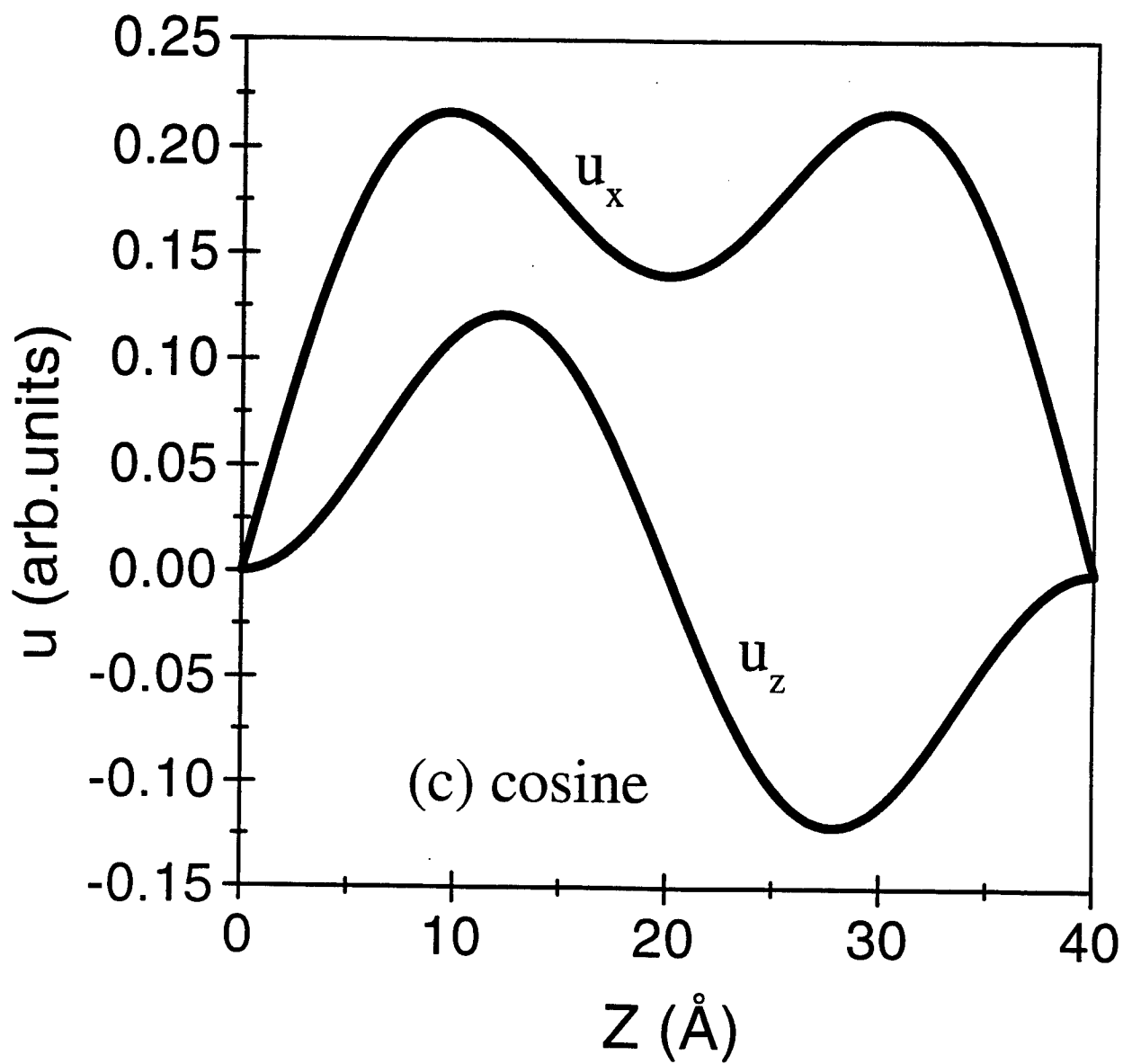
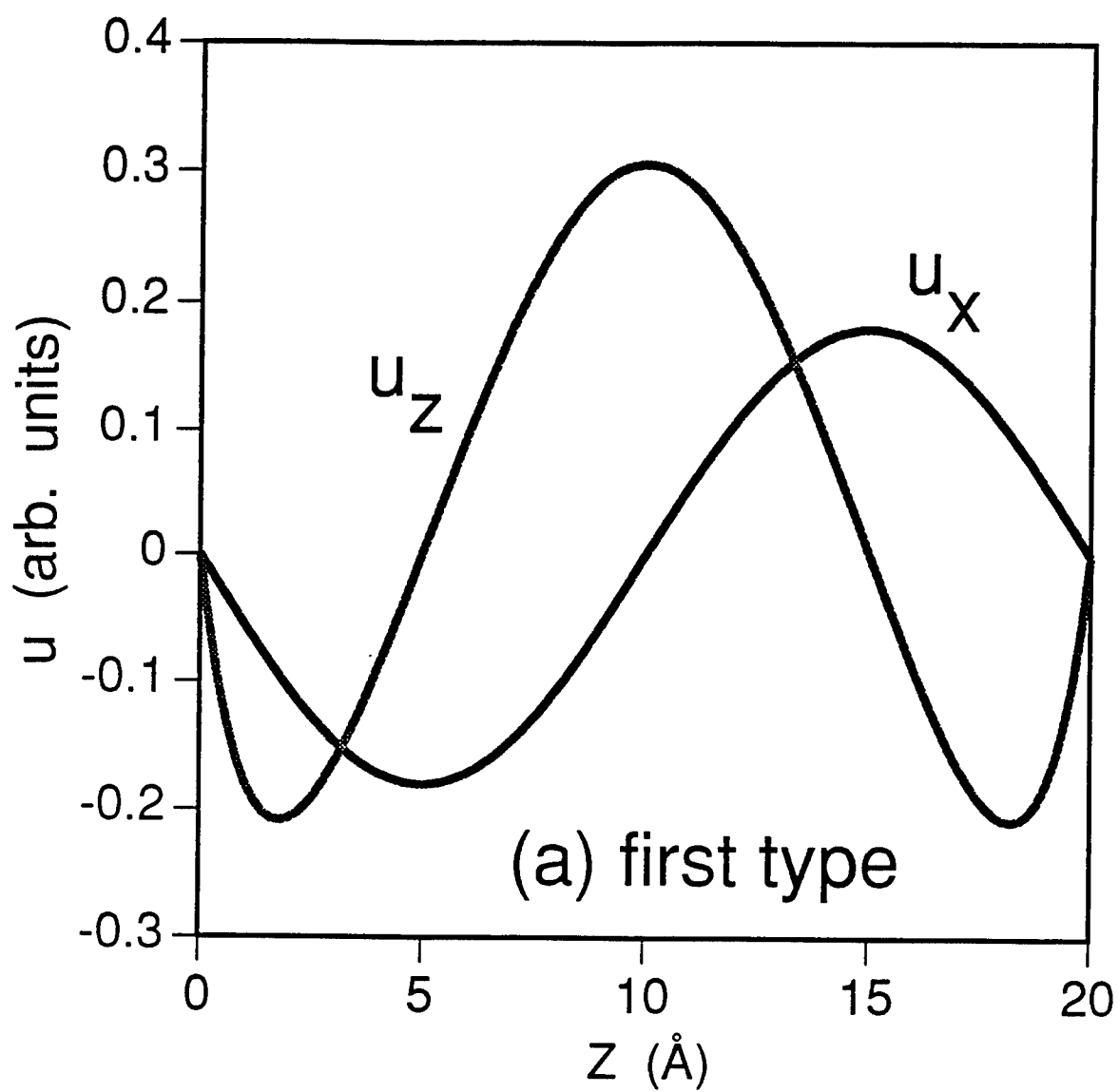


Fig.1 (c)



(a) first type

Fig. 2 (a)

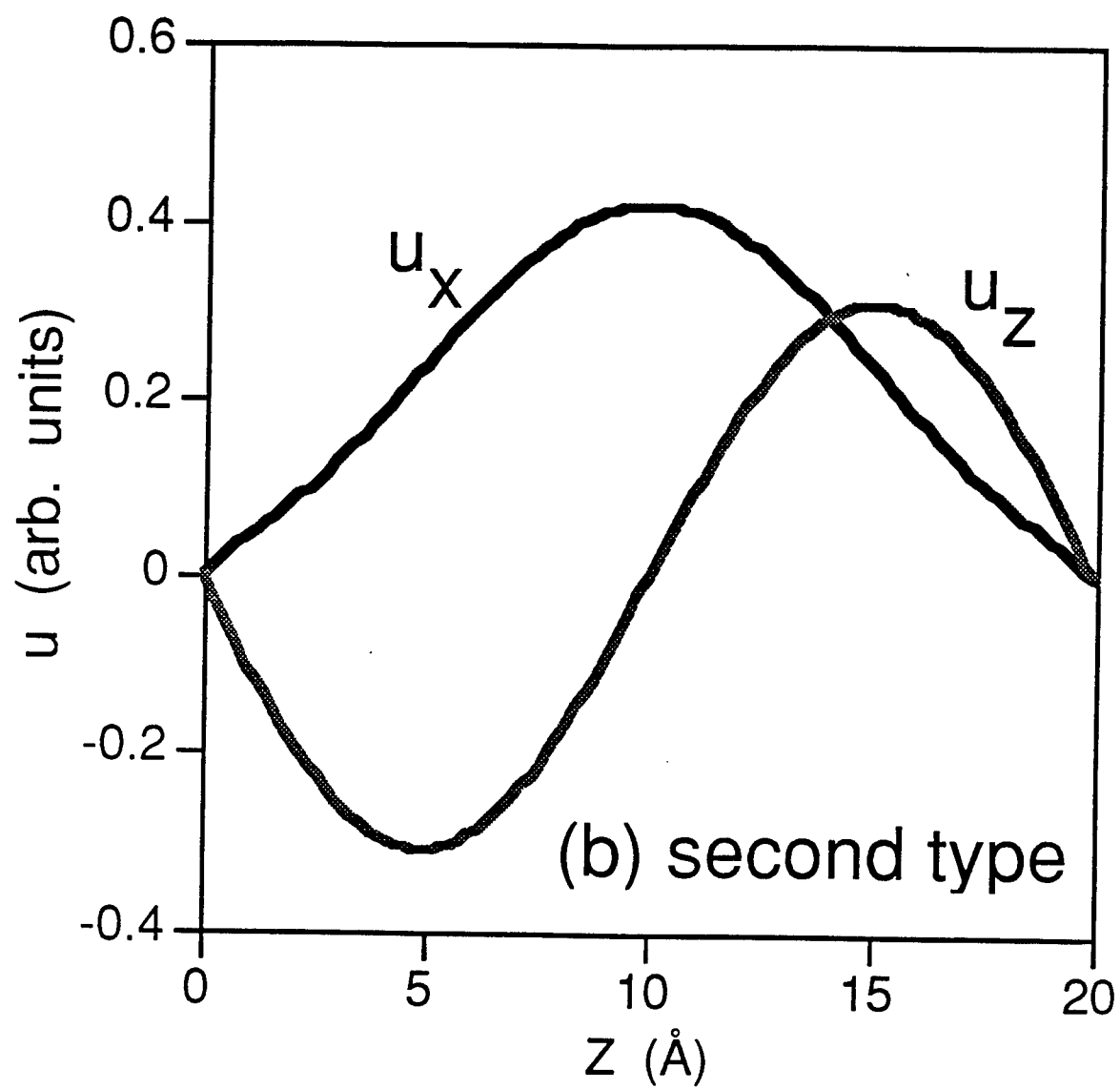


Fig. 2 (b)

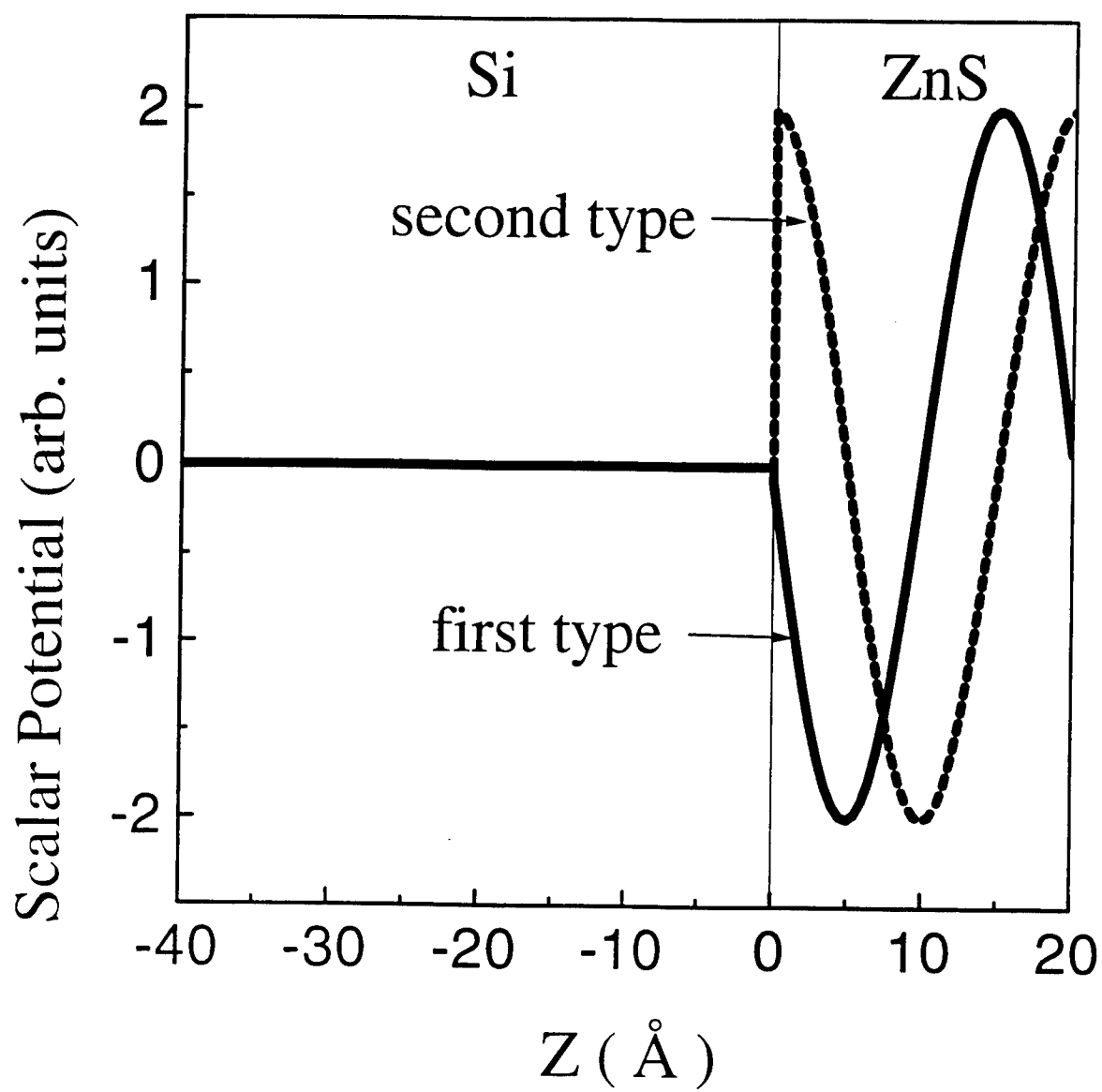


Fig. 3

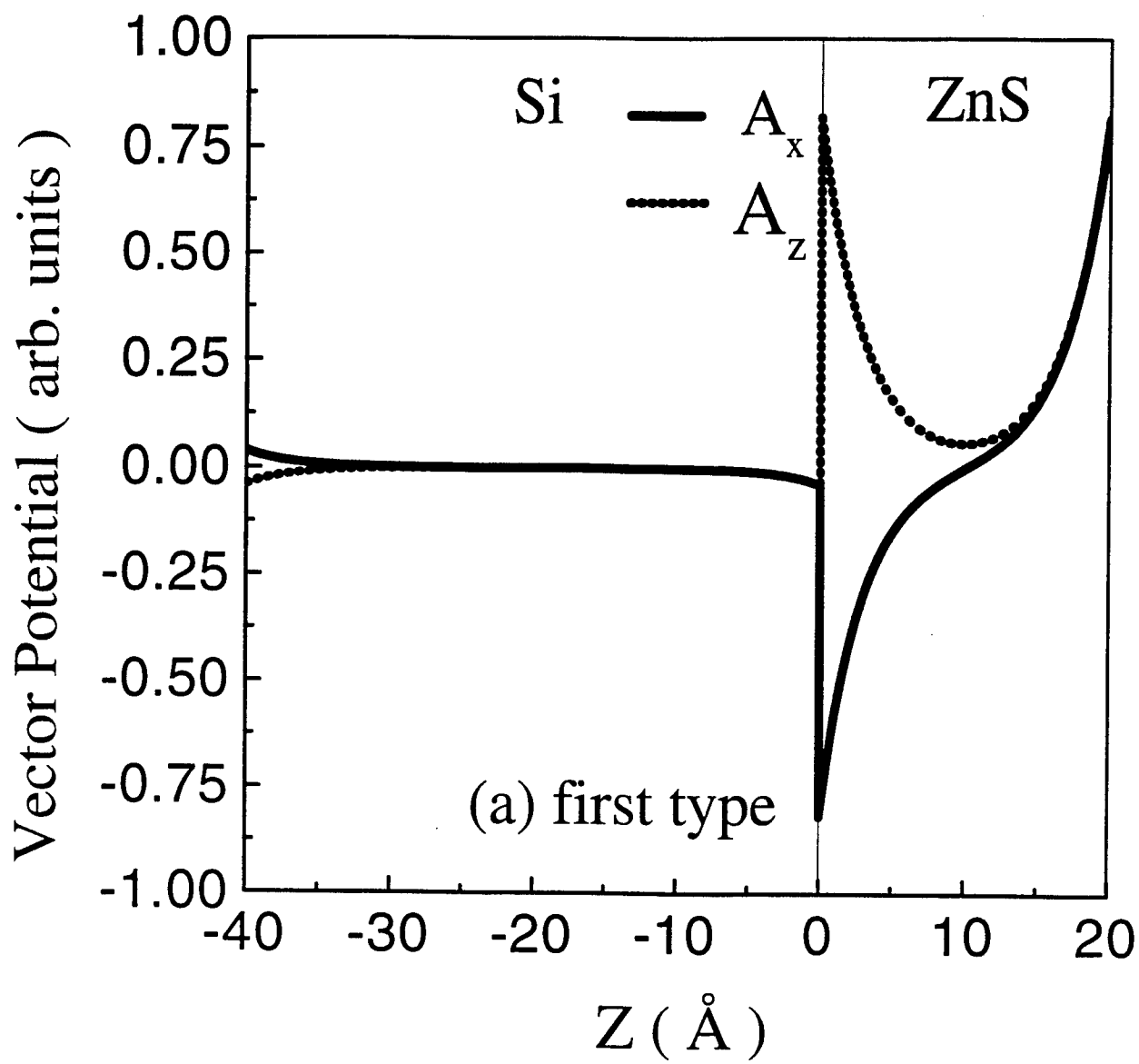


Fig. 4 (a)

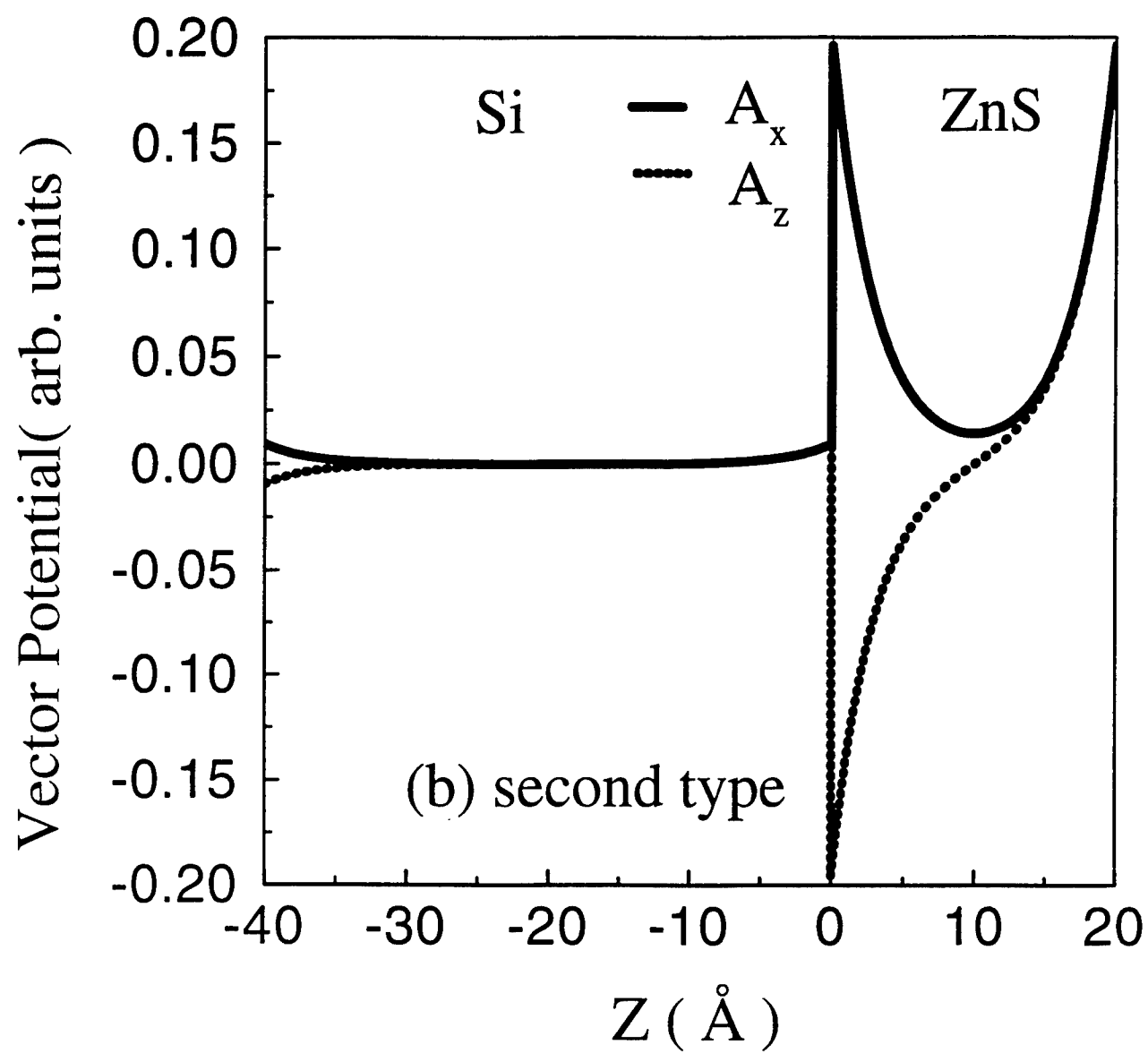


Fig. 4 (b)

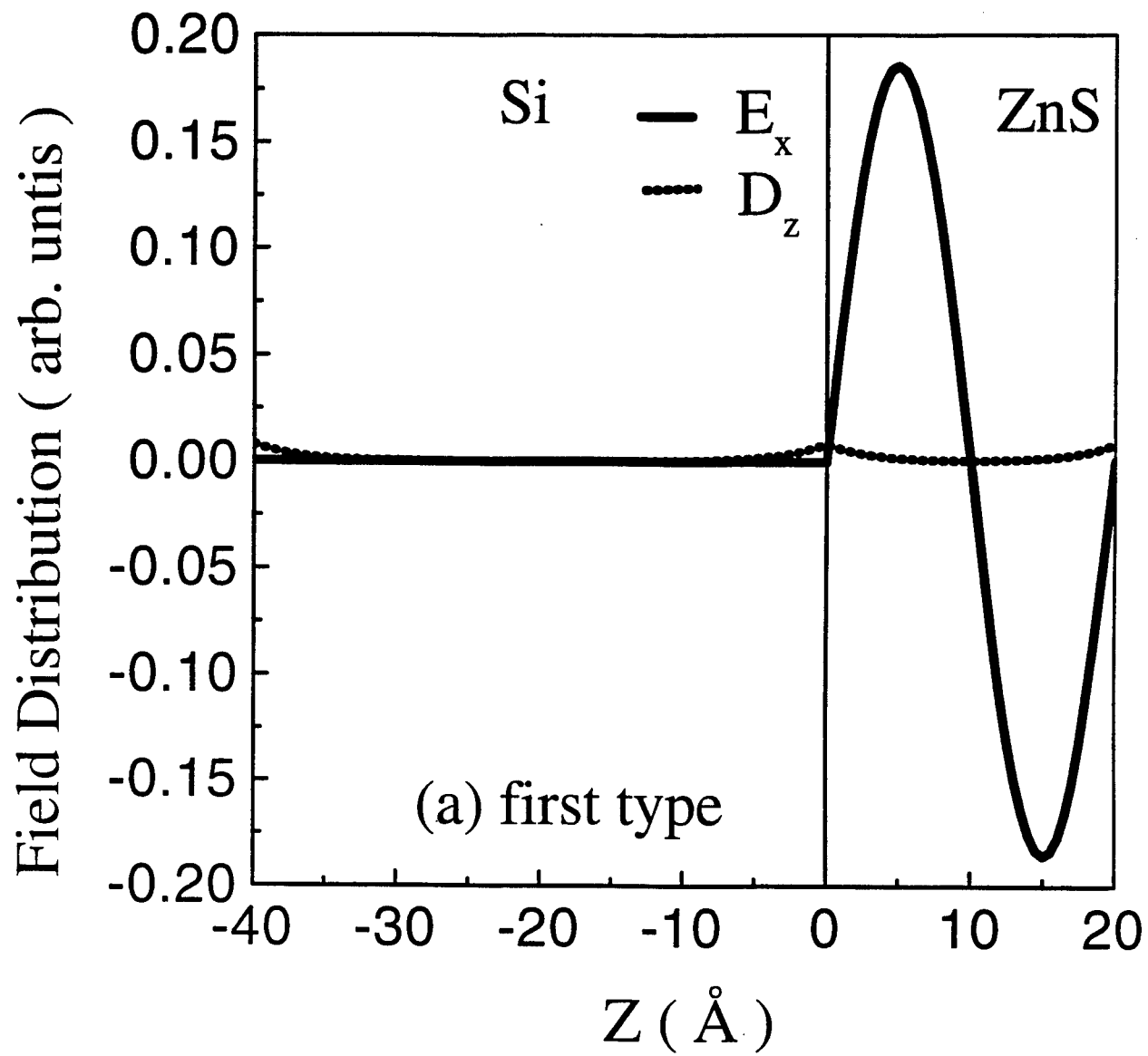


Fig. 5 (a)

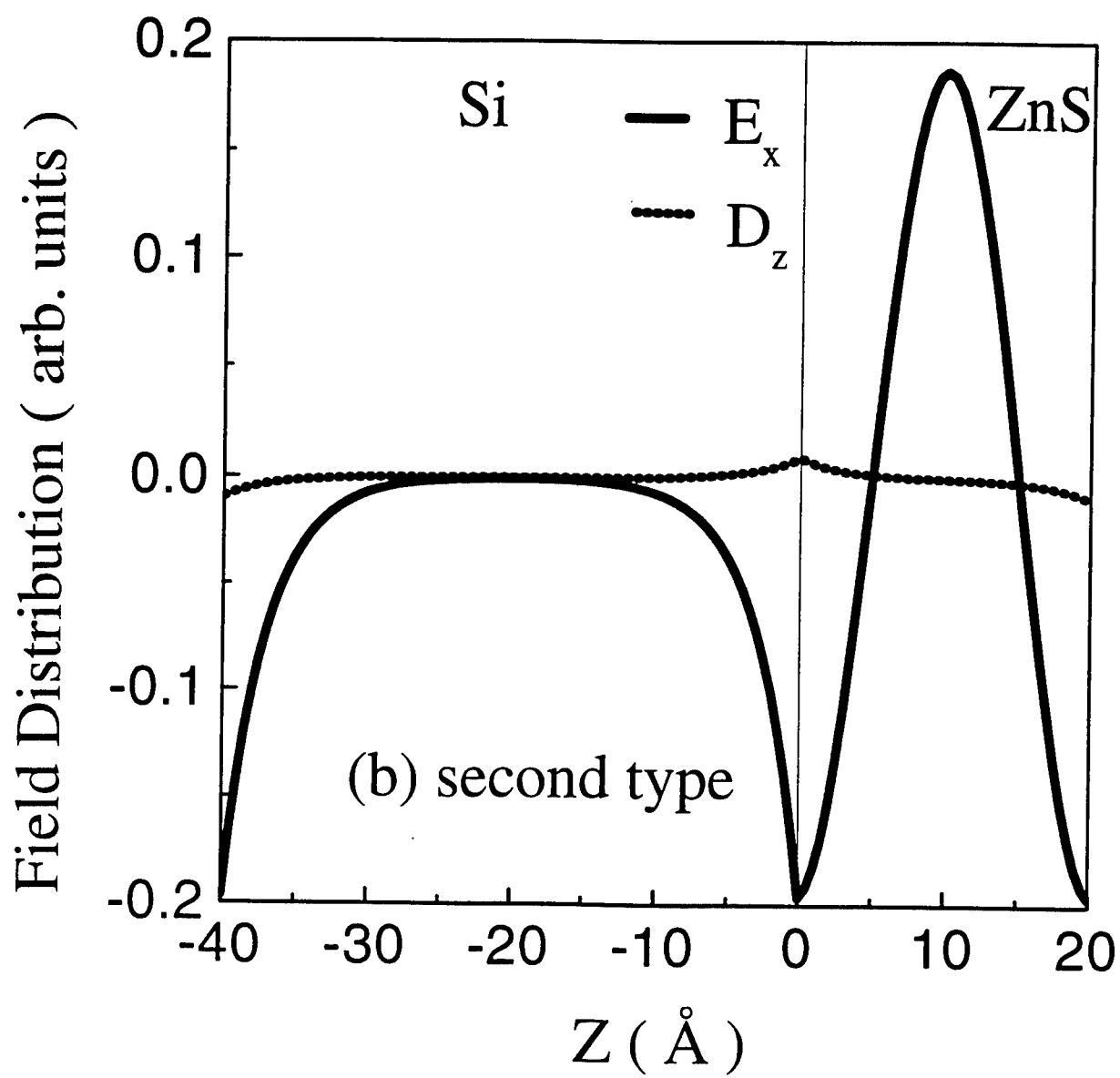


Fig. 5 (b)

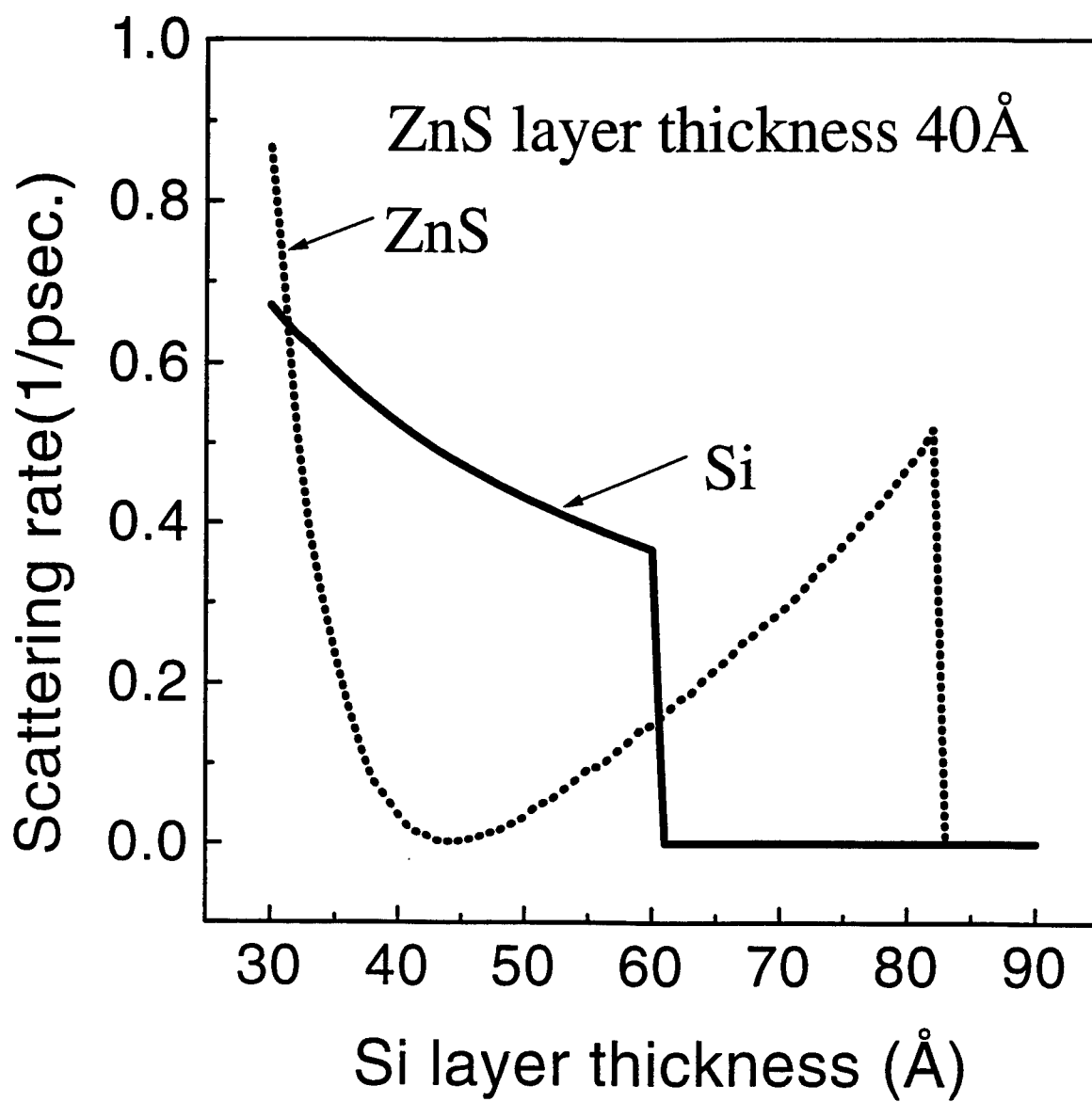


Fig. 6

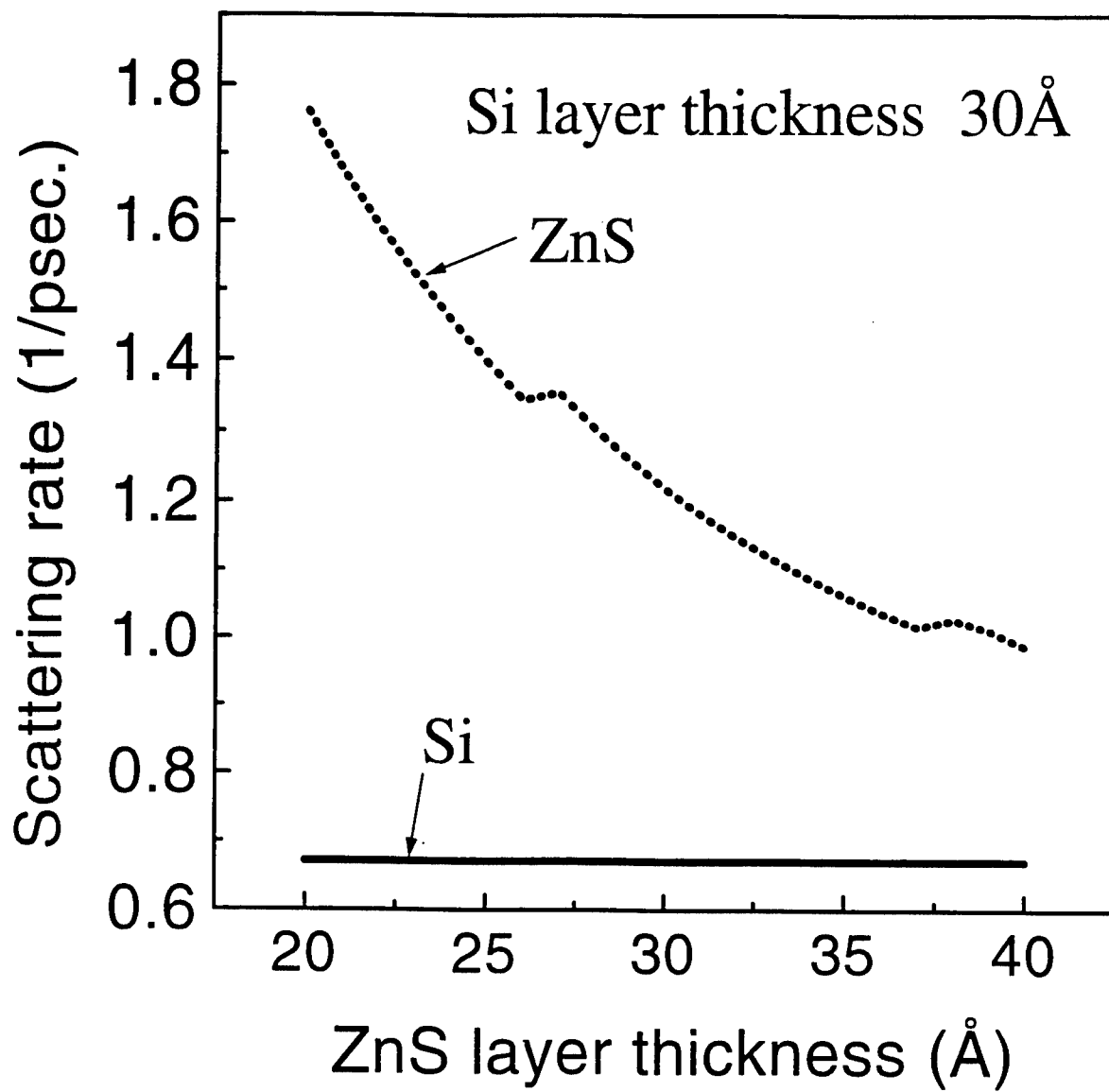


Fig. 7

A HARDWARE TEST-BED FOR ARRAY PROCESSING ALGORITHMS

Parker E. C. Bradley

Department of Mechanical, Aerospace & Manufacturing Engineering
Syracuse University
Syracuse, NY 13244
parker@pecb.com

Alan R. Lindsey, PhD, PE
Air Force Research Lab / IFGC
Rome, NY 13441

Dr. Donald Weiner
Dept. of Electrical Engineering
Syracuse University
Syracuse, NY 13244

Final Report for:
SREP Graduate Student Research Program
Air Force Research Lab / IFGC
Rome, NY

Sponsored by:
Air Force Office of Scientific Research
Bolling Air Force Base, DC
and
Air Force Research Lab / IFGC

December 1997

A HARDWARE TEST-BED FOR ARRAY PROCESSING ALGORITHMS

Parker E. C. Bradley

Department of Mechanical
& Manufacturing Engineering
Syracuse University
Syracuse, NY 13244
(parker@pecb.com)

ABSTRACT

This paper describes the design and use of a hardware and software environment for investigations in array processing (AP). The system unites the numerical processing abilities of MatLab with the real world through multichannel data acquisition hardware. Utilizing MatLab's scripting language, MEX-files, and C++ code modules, an adaptable, unified, and expandable software environment has been developed which can be readily modified for the equipment at hand. This modular software/hardware environment not only improves the ability of researchers to refine AP techniques, but assists in the integration of these tools into other technologies.

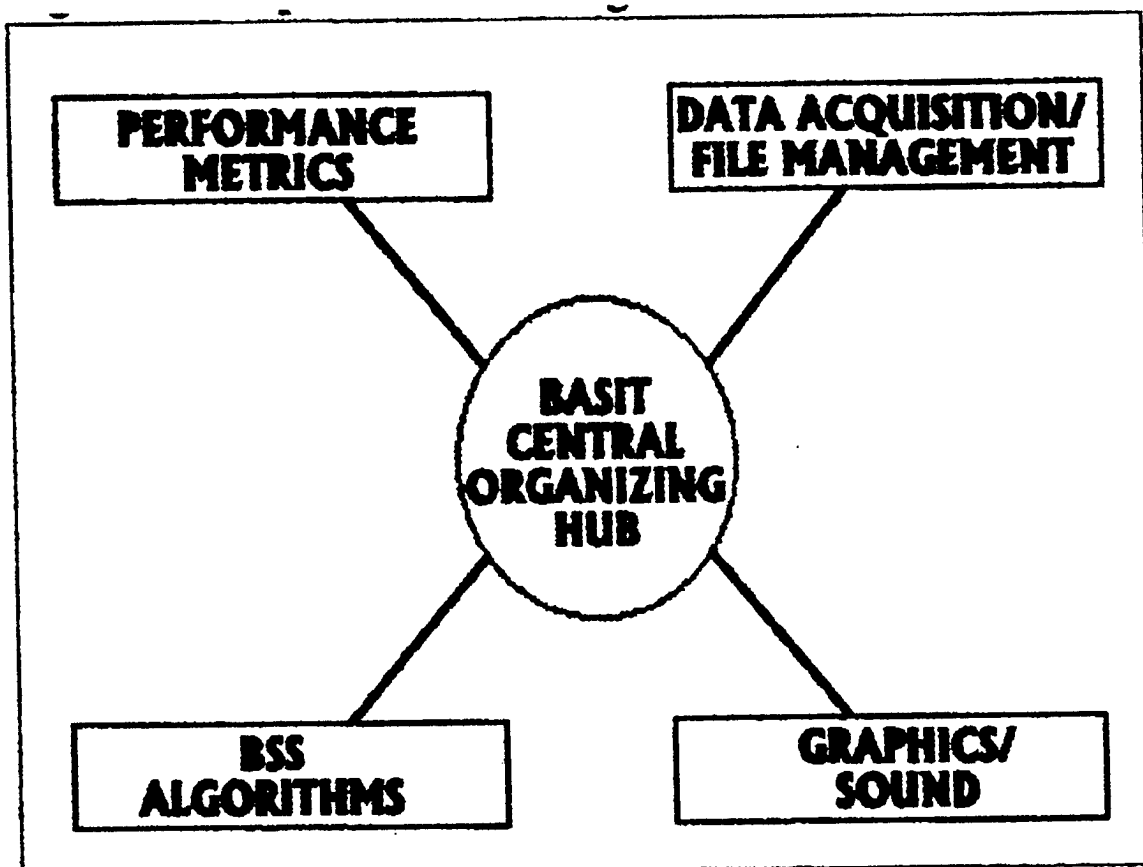


Figure 1: Simplified Block Diagram of AHTAPA Structure.

1. INTRODUCTION

The field of array processing encompasses a large number of algorithms and applications. To begin the task of developing a general purpose hardware test-bed for experimentation in this field, it is appropriate to narrow the focus to some subclass of algorithms with interesting properties. In the past two decades, a plethora of techniques have been spawned for adapting filters, demodulators, and receiver arrays to extract unknown signals of interest from corrupted data signals. These techniques have come to be referred to as *blind source separation (BSS)*. Examples of its main categories are:

- * *Coherence exploitation techniques*: utilizes higher-order spectral analysis and cummulants, and self-coherence restoration algorithms employing known spatial spectral, or self-coherence properties of the transmitted signal [1,2,3,4].
- * *Modulus restoration techniques*: utilize known modulus properties of the transmitted signal [5,6].
- * *Anticipatory techniques*: utilize maximum likelihood and burst-acquisition methods, which exploit

known temporal support of the transmitted signal [7].

The diversity of approaches to blind source separation leads to discussions of the relative merits of different approaches. Each approach has circumstances to which it is better suited than another approach, and sometimes a combination of techniques might be in order. The “best” algorithm for a given situation can be difficult to determine. Also, the development and testing of AP algorithms in a real-life scenario can be very difficult, and time consuming as tools are reinvented. This underscores the need for a unified, adaptable computer environment utilizing all the tools and information at hand —allowing rapid implementation and analysis of any AP algorithm. This paper presents such a computer environment.

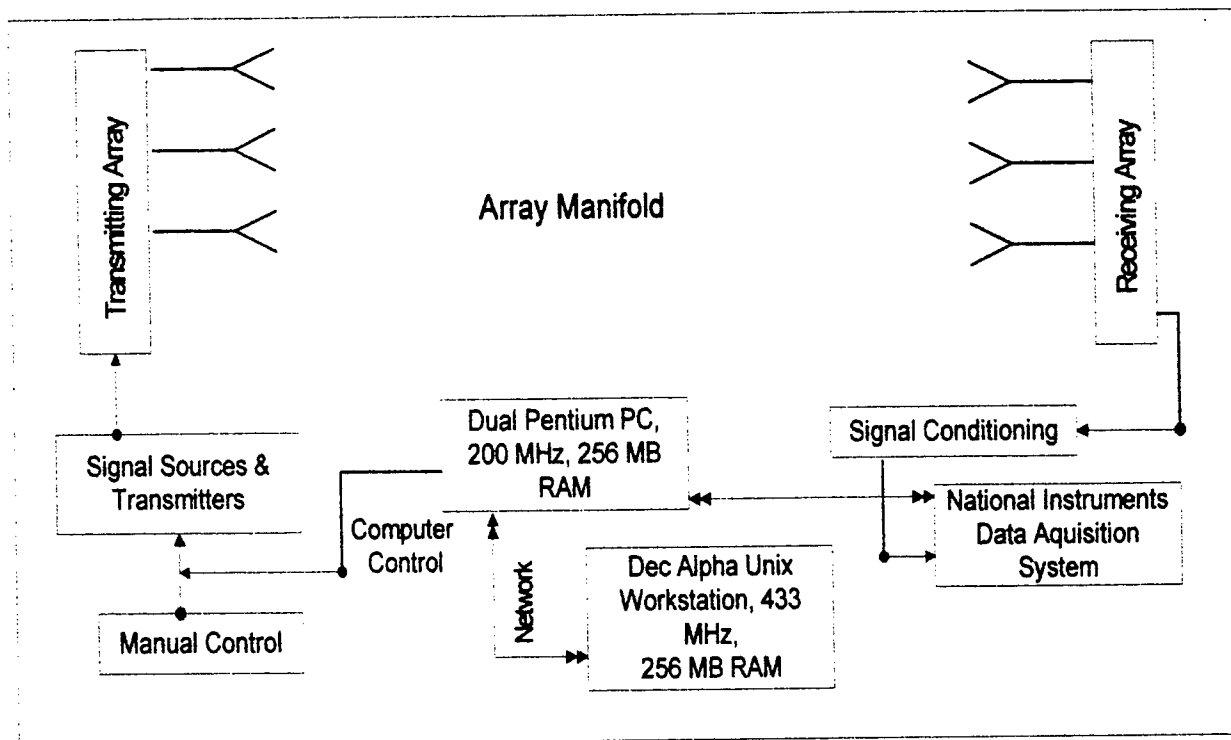
2. AHTAPA: UNDERLYING ARCHITECTURE

An environment fulfilling the above mentioned criteria is depicted in figures 1 and 2. The setup consists of a host workstation which provides overall control, multichannel data acquisition cards which collect data from multiple transducers (microphones, antennae, or light sensors, etc.), a Pentium Based PC for data collection and control, and assorted peripheral devices that facilitate input, output, control, and signal conditioning.

The host program provides an interface to the array processing modules, to data analysis and file management tools, and to other utilities that may be needed – thus unifying the entire research and development process into a single application. A simplified block diagram illustrating AHTAPA’s modularity is shown in Figure 1.

AHTAPA is a MatLab script which interacts with other MatLab scripts and C++ subroutines. In essence, the MatLab host script spawns a user interface serving as a front end and organizer to the other programming modules. New modules can be added in a simple “plug-n-play” fashion. Keeping in mind the goal of unifying the research and development process, a key aspect of the project is the continued development of a library of categorized array processing algorithms which can be expanded upon as needed, for comparison and contrast, and analyzed with a library of graphical, and analytic tools.

Figure2: Overview of array processing environment., and block diagram of Hardware



3. ARRAY HARDWARE

The collection of real world data is dependent upon the user's situation and resources. For example, during the first stage of this project's development, an external analog to digital converter (the WaveBook 512, manufactured by Iotech) with digital I/O was utilized. The unit has eight differential inputs and a maximum sampling rate of 1 MS/s. The Wavebook 512 was connected to an array of microphones, each of which had its own amplifier. The digitized data was then sent to a laptop where file handling and analysis was conducted through AHTAPA.

Eventually a dual-processor Pentium based PC was dedicated to data acquisition and control. The system utilizes two PCI-based data acquisition cards produced by National Instruments (PCI-MIO Series). Each card has a maximum sampling rate of 1 MHz at a 12 bit resolution, and eight differential channels. The computer has enough memory to acquire over 10 minutes of uninterrupted data from eight channels simultaneously. Sample rate, number of activated channels, and data-resolution can be adjusted.

The array system has been designed to allow for both real-time signal processing of real data, and for off-line data

analysis and algorithm development. A number of subsystems are incorporated into the current environment and include an anechoic chamber, a host workstation, a signal generation and broadcasting system, a signal collection/processing system, and miscellaneous peripherals. Figure 2 gives a block-diagram overview of the current hardware components and their interconnections.

However, a sophisticated setup is not a requirement for AHTAPA. The modular construction of AHTAPA allows a user to easily adapt AHTAPA to individual needs and resources – linking unique code modules, from the simplest game port input setup to the most elaborate DSP acquisition card. The program also comes with code modules, and instructions that could be used by a researcher to utilize common input devices to acquire real world data (at minimal cost) – such as serial, game, and parallel ports.

4. CONTROL SOFTWARE

With the current arrangement, the host machine running AHTAPA initiates data collection and control on the PC over the network, using a well established client-server model. A server running on the PC responds to requests issued by a client program running on the host workstation. The two exchange instructions, and the collected data is shared via an NFS network drive. While any program supporting sockets can do this MatLab makes an excellent client.

5. DATA ANALYSIS ISSUES

The successful application of any AP algorithm requires an understanding of the system from which the data was collected, as well as how it was collected. For example, intuition might lead to the assumption that the sound received by a microphone from multiple sources is a simple additive mixture. However, this is not the case. The sound “heard” at a point from multiple sources is a convoluted mixture. Another example is signals in the electromagnetic spectrum, be it radio, X-rays, or visible light. Intuition gives a fairly accurate picture in this case—that electromagnetic phenomena form a simple additive mixture. But there are exceptions to this as well, such as radar pulses through turbulent water, or visible images projected through glass with local impurities and an uneven/warped surface.

The manner in which the data is collected is also important. Under some circumstances the data collection process can introduce a convolution where there normally would not be one. For example: when a geological survey of the ocean bottom is conducted, sonar equipment is sometimes dragged along with the ship, this can generate turbulence around the sonar array that may introduce an additional convolution to the data being collected. Also, a particular AP algorithm may be sensitive to array geometry. A simple, generalized example of this problem is illustrated as follows.

Suppose there are two sources emitting a signal with wavelength x . If the sensors receiving the signal are a distance $0.03x$ apart, they could be behaving as a “single” sensor. The needed information to separate the signals from the mixture would then be lost.

The above are just a few examples of the many issues that must be taken into consideration when collecting data for AP analysis— and in deciding upon a particular AP algorithm for a given situation. As with any analytic tool, be it Fourier analysis, Wavelets, or Time Frequency Domain, knowledge of a method’s strengths and weaknesses is important, so that easy problems are not made difficult and mistakes not misunderstood as solutions.

6. SIMULATIONS

The AHTAPA environment allows for easy experimentation with algorithms, for example here we present a simulation that used an algorithm which was a combination of approaches from Belouchrani, Cardoso, neural networkd, and some trial and error experimentation. The simulation is performed using measured source signals, mixed with a realistic mixing matrix, estimated from experiments.

Using this new hybrid algorithm, with a constant learning rate, some interesting results were obtained and are depicted in figures 3,4,5,6, and 7.

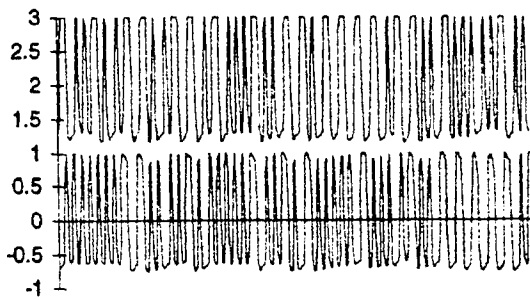


Figure 4. The observed signals $e1$ and $e2+4$ after mixing

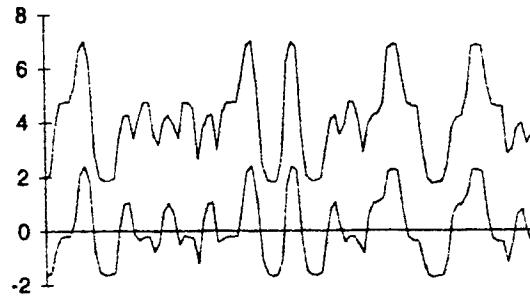


Figure 3 The pre-processed source signals x_i , and $x2+2$

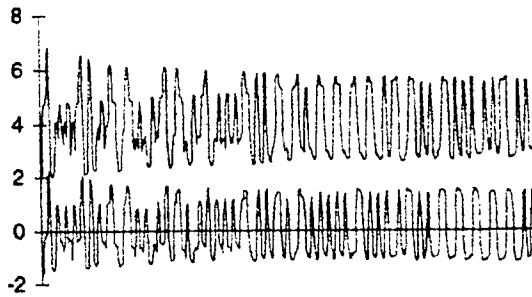


Figure 5. Output signals $s1$ and $s2+4$ during separation. **Figure 6.** Convergence of $c12$ and $c21$. The correct Solutions are the horizontal lines

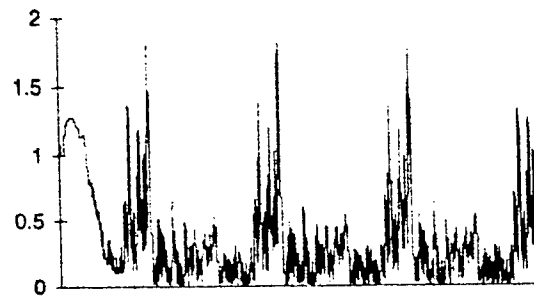
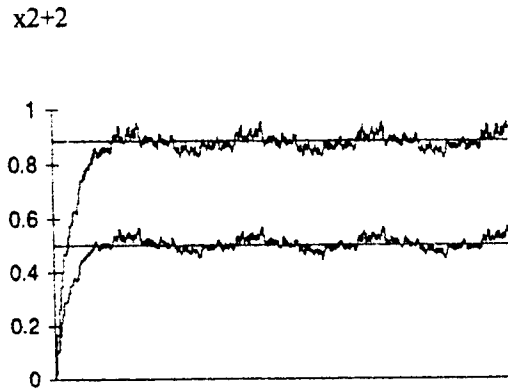


Figure 7. Gap values

From these results we can see that the coefficients converge rapidly toward their correct values; the output signals gradually regain the source signal shape; the *gap* is decreasing, but a slow oscillation, with about four periods is clearly visible in the graph; this oscillation can be seen in the convergence patterns of the coefficients.

This oscillation might be due to a dependence between the source signals x_i . Some of the causes for this coupling might be due to the following –Electromagnetic coupling between receivers; oscillation caused by small difference in frequency of the transmitted signals – which causes the patterns of both signals to slowly shift with respect to each other.

Since this hybrid algorithm determines the most independent components in a set of signals, we expect that the algorithm must be able to find this dependence. To verify this, the experiment was repeated using a unitary mixing matrix, so that the mix equals the source signals. If the signals are fully independent, no action should occur from the network of the algorithm. However, a dependence is shown in figure 8., where a fluctuation is seen of the coefficient around zero. The resulting output signals are the most independent components of the original source signals x_i .

Repeating the separation for mixed signals, using the most independent components, one would expect the convergence of the coefficients to be less distorted by the low frequency oscillation. This is verified in figure 9.

Figure 8. Convergence of c_{12} and $c_{21}+2$ for the

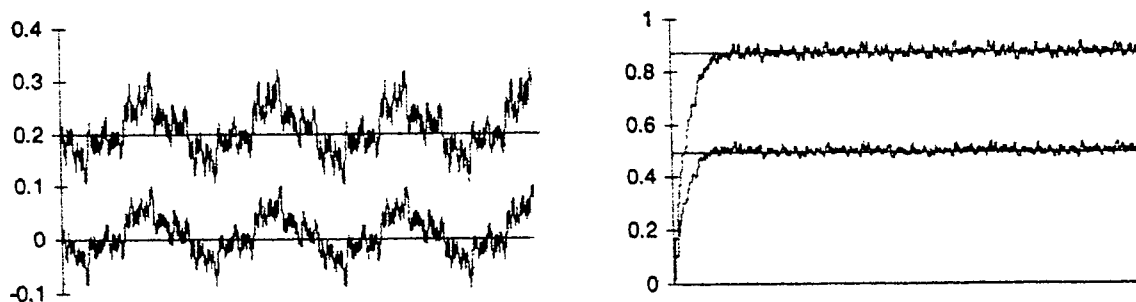


Figure 9. Convergence of c_{12} and c_{21} for the most unmixed source signals x_i

Independent components of the source signals x_i .

7. CONCLUSION

An adaptable, modular environment for experimentation in AP algorithms, based on MatLab scripts and C++ code is presented. The environment developed can be utilized by any researcher, and is adaptable to needs, and resources. The goal is that the work presented here may be developed into a standard tool for AP work—streamlining the development and testing process as well as assisting in integrating AP technologies into other disciplines.

References

1. Chen, Ling, Kusaka, Hiroji, Kominami, Masabumi, and Yin Qingyie, "Blind Identification of Noncausal AR models based on Higher-Order Statistics" in *Signal Processing*, Vol. 48, 1996, pg 27-36.
2. Nikias, Chrysostomos L., and Raghunveer Mysore R, "Bispectrum Estimation: A Digital Signal Processing Framework," in *Proc. IEEE*, Vol. 75, no. 7, July 1987, pp. 869-891.
3. Mendel, Jerry M. "Tutorial on Higher-Order Statistics (Spectra) in Signal Processing and System Theory: Theoretical Results and Some Applications," in *Proc. IEEE*, Vol 79, no.3, March 1991, pp. 279-305.
4. Agee, B.G., Schell, S.V., and Gardner, W.A., "Spectral Self-Coherence Restoral: A New Approach to Blind Adaptive Signal Extraction Using Antenna Arrays," in *Proc. IEEE*, Vol.78, no.4, April 1990, pp. 753-767.
5. Treichler, J.R., Larimore, M.L., "New Processing Techniques Based on the Constant Modulus Algorithm," in *IEEE Trans. ASSP*, April 1985, pp. 420-431.
6. Van der Veen, Alle-Jan, and Paulraj, Arogyaswami, "An Analytical Constant Modulus Algorithm," in *IEEE Trans. On Signal Processing*, Vol. 44, no.5, May 1996, pp 1136-1155.
7. Agee, B.G., "Fast Acquisition of Burst and Transient Signals Using a Predictive Adaptive Beam-Former," in *Proc. 1989 IEEE Military Communications Conf.*
8. Belouchrani, A., Cardoso, J.F., and Moulines, E., "Second Order Blind Separation of Temporally Correlated Sources." A paper received directly from author – further publication information is not known.